SPATIOTEMPORAL MODELING OF AIR POLLUTANTS AND THEIR HEALTH EFFECTS IN THE PITTSBURGH REGION

by

Tao Xue

Bachelor of Science, Peking University, China, 2011

Submitted to the Graduate Faculty of

the Department of Environmental and Occupational Health

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

TAO XUE

It was defended on

January 16 , 2015

and approved by

Dissertation Advisor:
Aaron Barchowsky, Ph.D
Professor
Department of Environmental and Occupational Health
Graduate School of Public Health
University of Pittsburgh

Committee Members:
Richard Bilonick, Ph.D
Assistant Professor
Department of Ophthalmology
School of Medicine
University of Pittsburgh

Jim Peterson, Ph.D
Associate Professor
Department of Environmental and Occupational Health
Graduate School of Public Health
University of Pittsburgh

Ravi Sharma, Ph.D
Assistant Professor
Department of Behavioral and Community Health Sciences
Graduate School of Public Health
University of Pittsburgh

Evelyn Talbott, Dr.P.H
Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Abdus Wahed, Ph.D
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Aaron Barchowsky, Ph.D

SPATIOTEMPORAL MODELING OF AIR POLLUTANTS AND THEIR HEALTH EFFECTS IN THE PITTSBURGH REGION

Tao Xue, PhD

University of Pittsburgh, 2015

ABSTRACT

Air pollutants have been associated with adverse health outcomes such as cardiovascular and respiratory diseases through epidemiological studies. Spatiotemporal and spatial statistics are widely used in both exposure assessment and health risk estimation of air pollutants. In the current paper, spatiotemporal and spatial models are developed for and applied to four specific topics about air pollutants: (1) estimating spatiotemporal variations of particulate matter with diameter less than 2.5 μm ($PM_{2.5}$) using monitoring data and satellite aerosol optical depth (AOD) measurements, (2) estimating long-term spatial variations of ozone (O_3) using monitoring data and satellite O_3 profile measurements, (3) spatiotemporal associating acute exposure of air pollutants to mortality, and (4) spatiotemporal associating chronic air pollution exposure to lung cancer incidence. Environmental, socioeconomic and health data from Allegheny county and the State of Pennsylvania are collected to illustrate these techniques.

The public health significance of these studies includes characterizing the exposure level of air pollutants and their health risks for mortality caused by cardiovascular and respiratory diseases and lung cancer incidence in the Pittsburgh region and developing novel spatiotemporal models such as spatiotemporal generalized estimating equations for the regression analysis of spatiotemporal counts data, especially for the massive spatiotemporal data used in epidemiological studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Air pollutants such as particulate matter with aerodynamic diameter $\leq 10\mu m$ (PM_{10}) and $\leq 2.5\mu m$ ($PM_{2.5}$), Ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO) have been associated with adverse health effects including cardiovascular and respiratory diseases [Dominici et al., 2006, Peng et al., 2009, Glad et al., 2012], infant birth defects [Ritz et al., 2000, Salam et al., 2005, Sapkota et al., 2012], DNA damage [Wei et al., 2009, Ren et al., 2010], cancer mortality [Laden et al., 2006, Pope et al., 2011], and many others. However, most of the epidemiological studies of air pollutants are longitudinal studies for city or metropolitan areas [Pope III et al., 1991, Pope III and Dockery, 1992, Roemer et al., 1993, Rich et al., 2012, Gan et al., 2013] or cross-sectional ones for subnational or national regions [Dockery et al., 1989, Dijkema et al., 2011, Correia et al., 2013]. In these longitudinal studies, spatial variations of air pollutants were usually ignored due to difficulties in accurately characterizing spatial or spatiotemporal variations of exposure to air pollutants at the city or metropolitan level and in estimating health effects with complex spatiotemporal autocorrelation in regression models. However, previous studies have shown that spatial variations of air pollutants were comparable to their temporal variations especially for gaseous pollutants (e.g. O_3 and SO_2) [Wade et al., 2006], so that ignoring spatial variations in longitudinal studies may cause potential exposure misclassification. While in national or subnational scale cross-sectional studies researchers usually capture spatial variations of air pollutants between different cities but ignore the variations within cities, but recent research has shown that within-city health effects of air pollutants may be larger than between-city health effects [Jerrett et al., 2005]. Therefore, spatial or spatiotemporal analysis may be critical for both exposure assessment and health risk estimation of air pollutants.

The current paper is focused on the spatiotemporal analysis of metropolitan scale epidemiology of air pollutants in order to reduce exposure misclassification and to increase accuracy in health risk estimation. In contrast to the traditional longitudinal study of health effects of air pollutants, the spatiotemporal study will consider spatial variations of air pollutants and health outcomes simultaneously with their temporal variations. Specifically, this study aims to (1) **improve the assessment of spatiotemporal variations of air pollutants exposure** and (2) **develop novel methods to deal with spatiotemporal autocorrelation in associating health outcomes with air pollutants**. Spatiotemporal models, which play a key role in the two aims, are usually developed through extending traditional spatial statistics to spatiotemproal dimensions [Cressie and Wikle, 2011] and have been introduced into environmental [Guttorp et al., 1994, Wikle and Cressie, 1999, Sahu et al., 2007, Katzfuss and Cressie, 2011] and public health studies [Waller et al., 1997, Xia and Carlin, 1998, Mugglin et al., 2002, Schmid and Held, 2004] in recent years. In the current study, we will apply appropriate existing spatiotemporal models to and develop novel spatiotemporal models for the aforementioned aims.

To illustrate the spatiotemporal statistics in exposure assessment and health effects estimation of air pollutants, my thesis research consists of four specific topics:

1. Estimating the spatiotemporal variations of $PM_{2.5}$ mass concentrations using monitoring satellite aerosol optical depth (AOD) data over the Pittsburgh region, 2001-2008;
2. Estimating the spatial distribution of O_3 in the continental United States from ozone measurement instrument (OMI) O_3 profile using a latent vector spatial model;
3. Associating mortality counts to air pollutants from 1999 to 2009 in the Pittsburgh region using spatiotemporal generalized estimating equations;
4. Spatiotemporal associating lung cancer incidence to Ozone and smoking in the state of Pennsylvania, 2001-2007.

The first two topics are focused on exposure assessment of air pollutants using spatiotemporal statistics (Aim 1), while the last two are focused on regressing spatiotemporal health outcomes on air pollutants (Aim 2). In the four topics, we will illustrate how to apply a two-step generalized additive mixed model, a latent vector spatial model, spatiotemporal

generalized estimating equations and a Bayesian hierarchical model in air pollutants studies. Among these methods, the spatiotemporal generalized estimating equations is a novel method developed for regressing spatiotemporal counts data. The other three models are derived from existing methods. As most of the data that we used to illustrate these methods are from the Pittsburgh region, our study has characterized the exposure and health risks of air pollutants in Pittsburgh in detail.

The remaining sections will be organized as follows: In Chapter 2, we review the existing spatial and spatiotemporal models that are related to our studies. From Chapter 3 to Chapter 6, we are going to show the aforementioned four topics one by one. In Chapter 7 we summarize our findings about air pollutants' exposure and health risks in the Pittsburgh Region and illustrate the potential use of our statistical methods in further studies.

2.0 A BRIEF REVIEW OF SPATIAL AND SPATIOTEMPORAL STATISTICS

In this chapter, we are going to briefly review the basic statistical models that have been widely used in air pollution and public health studies, including spatial and spatiotemporal point process and lattice process. We will focus on two methods: spatiotemporal Kriging and the conditional autoregressive (CAR) model. The former has been widely used to interpolate spatiotemporal measurements of air pollutants, while the latter has been used widely in modeling areal measurements of health outcomes, e.g. counts of diseases by census tracts. Air monitoring data in the Pittsburgh region from 1999 to 2009 will be used to illustrate the former methods, while for the latter method, we are going to apply it in a study to associate cancer incidence ratios with socioeconomic factors in Allegheny County from 2000 to 2011.

2.1 SPATIAL AND SPATIOTEMPORAL POINT PROCESS

2.1.1 Spatial Point Process

Spatial point processes have been applied to model spatial measurements with exact geographic locations, e.g. air pollution monitoring data [Jun and Stein, 2004, Wong et al., 2004, Fuentes and Raftery, 2005]. Even through statisticians are more interested in non-stationary point process [Fuentes, 2001, Schmidt and O'Hagan, 2003], the most widely used spatial point model is the stationary multivariate Gaussian process [Cressie, 1993]:

$$\boldsymbol{z} \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } (\boldsymbol{\Sigma})_{i,j} = \sigma^2 \rho(||\boldsymbol{s}_i - \boldsymbol{s}_j||_2), \tag{2.1}$$

where $\boldsymbol{z}$ is a $n \times 1$ vector and denotes the spatial measurements, $\boldsymbol{s}_i$ denotes the spatial coordinates for i^{th} measurements and $\boldsymbol{\Sigma}$ is the $n \times n$ variance-covariance matrix for $\boldsymbol{z}$. In Equation 2.1, the expectation of $\boldsymbol{z}$ can be a constant μ, or be modified as a linear function of covariates ($\boldsymbol{\mu} = \boldsymbol{X\beta}$) or a smooth function of spatial coordinates ($\boldsymbol{\mu} = f(\boldsymbol{s})$); the variance-covariance matrix $\boldsymbol{\Sigma}$ is independent of specific location but instead is a function of the distance between two measurements, and known as the covariance function: $C(\cdot|\sigma^2, \boldsymbol{\theta}) \doteq \sigma^2 \rho(\cdot|\boldsymbol{\theta})$. The inferences for parameters in the covariance function can be made using a Bayesian Markov Chain Monte Carlo (MCMC) simulation or using method of moments estimator (the *varigram* approach in geo-statistics).

2.1.2 Variograms and Kriging

The concept of the variogram is closely related with the covariance and covariance function. The variogram (2γ) between two spatial measurements z_i and z_j is defined as:

$$2\gamma_{i,j} = E\left([(z_i - \mu_i) - (z_j - \mu_j)]^2\right). \tag{2.2}$$

Under the stationarity assumption, the expectation of $\boldsymbol{z}$ is a constant, thus the semi-variogram (γ) can be simplified as $\gamma_{i,j} = 1/2E\left((z_i - z_j)^2\right)$, which can be related with covariance function as follows:

$$\begin{aligned} \gamma_{i,j} &= 1/2\{E(z_i - z_j)^2 + Var(z_i - z_j)\} \\ &= 1/2\{Var(z_i) + Var(z_j) - 2Cov(z_i, z_j)\} \\ &= C(0) - C(||\boldsymbol{s_i} - \boldsymbol{s_j}||_2). \end{aligned}$$

Similar to the covariance function, variograms are functions of distances, and thus we can also denote $\gamma_{i,j}$ as $\gamma(||\boldsymbol{s_i} - \boldsymbol{s_j}||_2)$. Assuming that two spatial measurements between infinite distance are independent ($C(\infty) = 0$), we can derive a variogram of infinite distance via the above equation as $\gamma(\infty) = C(0)$, which is known as the *sill* of variogram. Therefore, the covariance can be expressed as:

$$C(||\boldsymbol{s_i} - \boldsymbol{s_j}||_2) = \gamma(\infty) - \gamma(||\boldsymbol{s_i} - \boldsymbol{s_j}||_2). \tag{2.3}$$

Inference about variograms can be made using point-wise moment estimators based on pairs of measurements between distances of a specific range:

$$\hat{\gamma}(d) = \frac{1}{2|N(d)|} \sum_{(i,j)\in N(d)} (z_i - z_j)^2, \text{ where } N(d) = \{(i,j)|\ \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \in [d-\delta, d+\delta)\},$$

which is known as empirical variogram. However, the variance-covariance matrix derived from the empirical variogram is not guaranteed to be positive definite. Therefore, we usually refit the empirical variograms using functions of specific forms, e.g. exponential, spherical and Matérn functions. For more details about variograms, see Cressie, (1993).

Kriging is a method to interpolate point values via weighted linear combination of existing spatial measurements. Depending on whether the expectation ($\boldsymbol{\mu}$) is zero, there are two specific methods, *simple Kriging* and *Ordinary Kriging*. In the rest of this section, we will review Kriging methods as an aspect of optimization to illustrate the core concepts of spatial interpolation.

Let $\boldsymbol{s}^*$ denote the coordinates of the interpolated points and $\boldsymbol{z}^* = \boldsymbol{W}\boldsymbol{z}$ denote the weighted linear combination of interpolated values, so that Kriging can be viewed as a optimization problem:

$$\begin{aligned} \underset{\boldsymbol{W}}{\text{argmin}} \quad & E\left((\boldsymbol{z}^* - \boldsymbol{W}\boldsymbol{z})^2\right) \\ \text{subject to} \quad & \boldsymbol{W}\mathbf{1} = \mathbf{1}, \text{ where } \mathbf{1}_{n\times 1} = [1, \cdots, 1]'. \end{aligned} \tag{2.4}$$

Under the stationarity assumption, the objective function in Equation 2.4 can be expressed by covariances:

$$\begin{aligned} E\left((\boldsymbol{z}^* - \boldsymbol{W}\boldsymbol{z})^2\right) &= Var(\boldsymbol{z}^* - \boldsymbol{W}\boldsymbol{z}) \\ &= Var(\boldsymbol{z}^*) + \boldsymbol{W}Var(\boldsymbol{z})\boldsymbol{W}' - 2\boldsymbol{W}Cov(\boldsymbol{z}^*, \boldsymbol{z}) \\ &= \text{Constant} + \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}' - 2\boldsymbol{C}\boldsymbol{W}', \end{aligned}$$

where $\boldsymbol{\Sigma} = Var(\boldsymbol{z})$ and $\boldsymbol{C} = Cov(\boldsymbol{z}^*, \boldsymbol{z})$, which can be estimated using the fitted variogram (or covariance function) as $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{C}}$. Therefore, Equation 2.4 can be rewritten as follows:

$$\begin{aligned} \underset{\boldsymbol{W}}{\text{argmin}} \quad & \boldsymbol{W}\hat{\boldsymbol{\Sigma}}\boldsymbol{W}' - 2\hat{\boldsymbol{C}}\boldsymbol{W}' \\ \text{subject to} \quad & \boldsymbol{W}\mathbf{1} = \mathbf{1}, \text{ where } \mathbf{1}_{n\times 1} = [1, \cdots, 1]', \end{aligned} \tag{2.5}$$

which is identical with *ordinary Kriging*. If the expectation $\boldsymbol{\mu}$ is zero, the constraint is unnecessary. Thus, regardless of the constraint, the optimized Kriging weights are $\hat{\boldsymbol{W}} = \hat{\boldsymbol{C}}(\hat{\boldsymbol{\Sigma}})^{-1}$, which is known as *simple Kriging*.

Under the optimization view of Kriging, we can naturally extend the interpolating method by adding further constrains. For example, we can restrict all the Kriging weights to be positive values, which may potentially improve Kriging performance in some scenarios (for details, see Page 143 in Cressie, (1993)).

2.1.3 Spatiotemporal Point Process and Kriging

Similar to the spatial process in Section 2.1.1, treating temporal coordinates as an additional dimension, we can define a stationary spatiotemporal point process as follows:

$$\boldsymbol{z} \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } (\boldsymbol{\Sigma})_{i,j} = \sigma^2 \rho_{st}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2, \|t_i - t_j\|_2), \tag{2.6}$$

where $\boldsymbol{z}$ is a $n \times 1$ vector and denotes spatiotemporal measurements and $\boldsymbol{\Sigma}$ is the $n \times n$ variance-covariance matrix for $\boldsymbol{z}$. Differing from Equation 2.1, a spatiotemporal point process captures both spatial and temporal dependence using a bivariate correlation function ρ_{st}. As the spatial and temporal dependence can be separably or non-separably mixed with each other [Cressie and Huang, 1999, Iaco et al., 2001, Gneiting, 2002, Bruno et al., 2009], spatiotemporal correlation functions can be defined by various forms. For example, one of the simplest separable spatiotemporal correlation functions is a linear combination of spatial and temporal correlation functions. As previous studies have shown that for air pollutants, spatiotemporal correlation functions are usually non-separable [Guttorp et al., 1994, De Iaco et al., 2002a, Bruno et al., 2009], in the following sections, we will consider a product-sum form of spatiotemporal dependence, which has been shown to be flexible and outperform other non-separable dependence structure [De Iaco, 2010]. A general form of the product-sum covariance function (C_{st}) can be defined as follows

$$\begin{aligned} C_{st}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2, \|t_i - t_j\|_2) = \; & k_* C_s(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2) + k_t C_t(\|t_i - t_j\|_2) \\ & + k_{st} C_s(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2) C_t(\|t_i - t_j\|_2), \end{aligned} \tag{2.7}$$

where C_s and C_t can be any existing forms of covariance function, e.g. *exponential*, *spherical* and *Matérn*. If the parameter k_{st} is zero, the product-sum covariance function will be reduced to separable.

Spatiotemporal variograms and spatiotemporal Kriging can be done analogously to Section 2.1.2 using a specific form of product-sum covariance function. However, considering the computing complexity for massive spatiotemporal data, we usually use neighboring measurements within a specific temporal period, which can be written as a optimization problem similar to Equation 2.4 as follows:

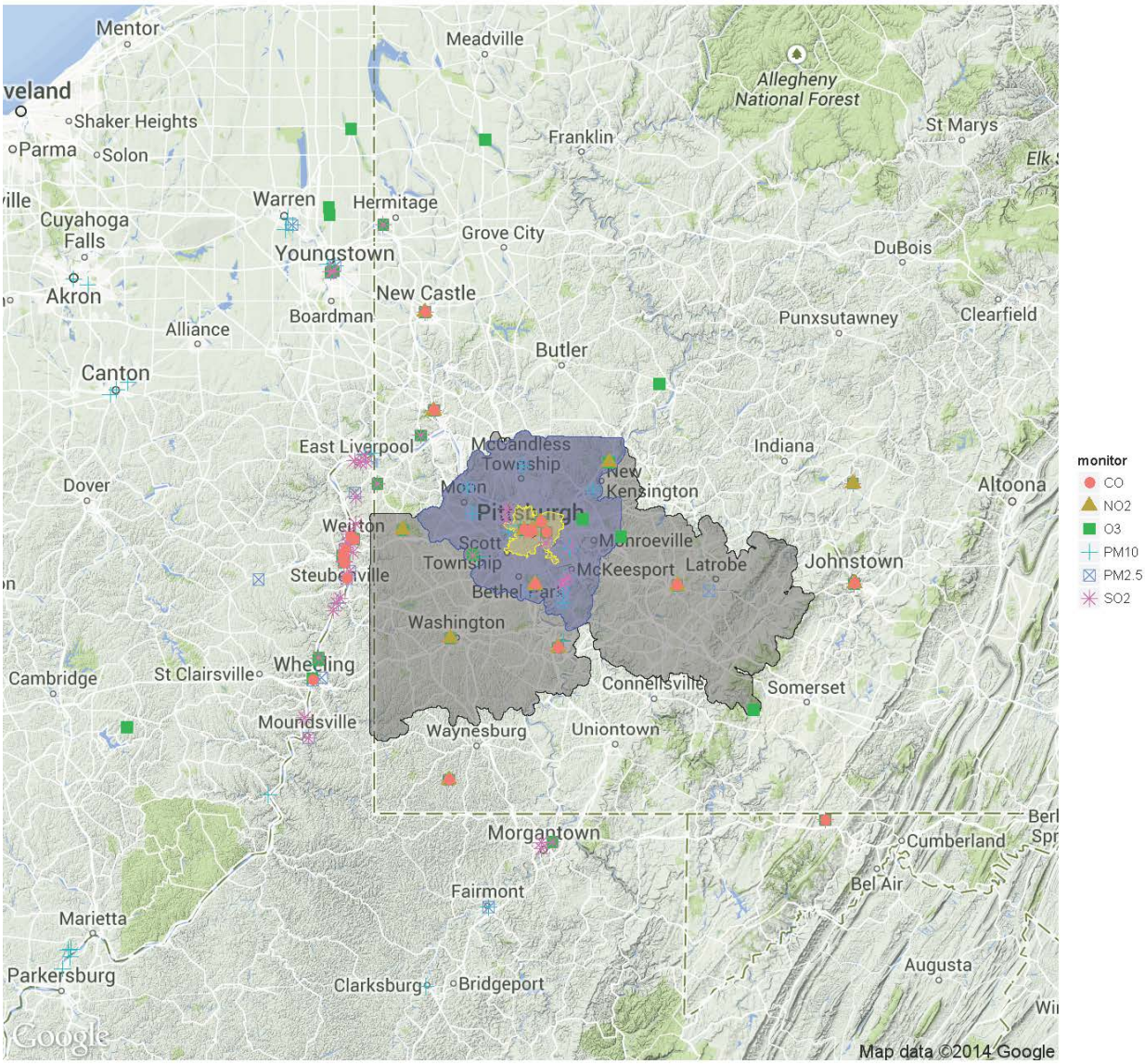
$$\begin{aligned} \underset{w_{s,t}}{\text{argmin}} \quad & E\left((z^*_{s',t'} - \sum_{s\in\mathcal{D},|t-t'|\leq\delta} w_{s,t} z_{s,t})^2\right) \\ \text{subject to} \quad & \sum_{s\in\mathcal{D},|t-t'|\leq\delta} w_{s,t} = 1, \end{aligned} \tag{2.8}$$

where (s,t) denotes the index of spatial and temporal measurements and $\mathcal{D}$ is a set of all spatial points in the study domain.

2.1.4 An Example of Spatiotemporal Kriging: Estimating Spatiotemporal variations of Air Pollutants Using Monitoring Data in the Pittsburgh Region, 1999-2009

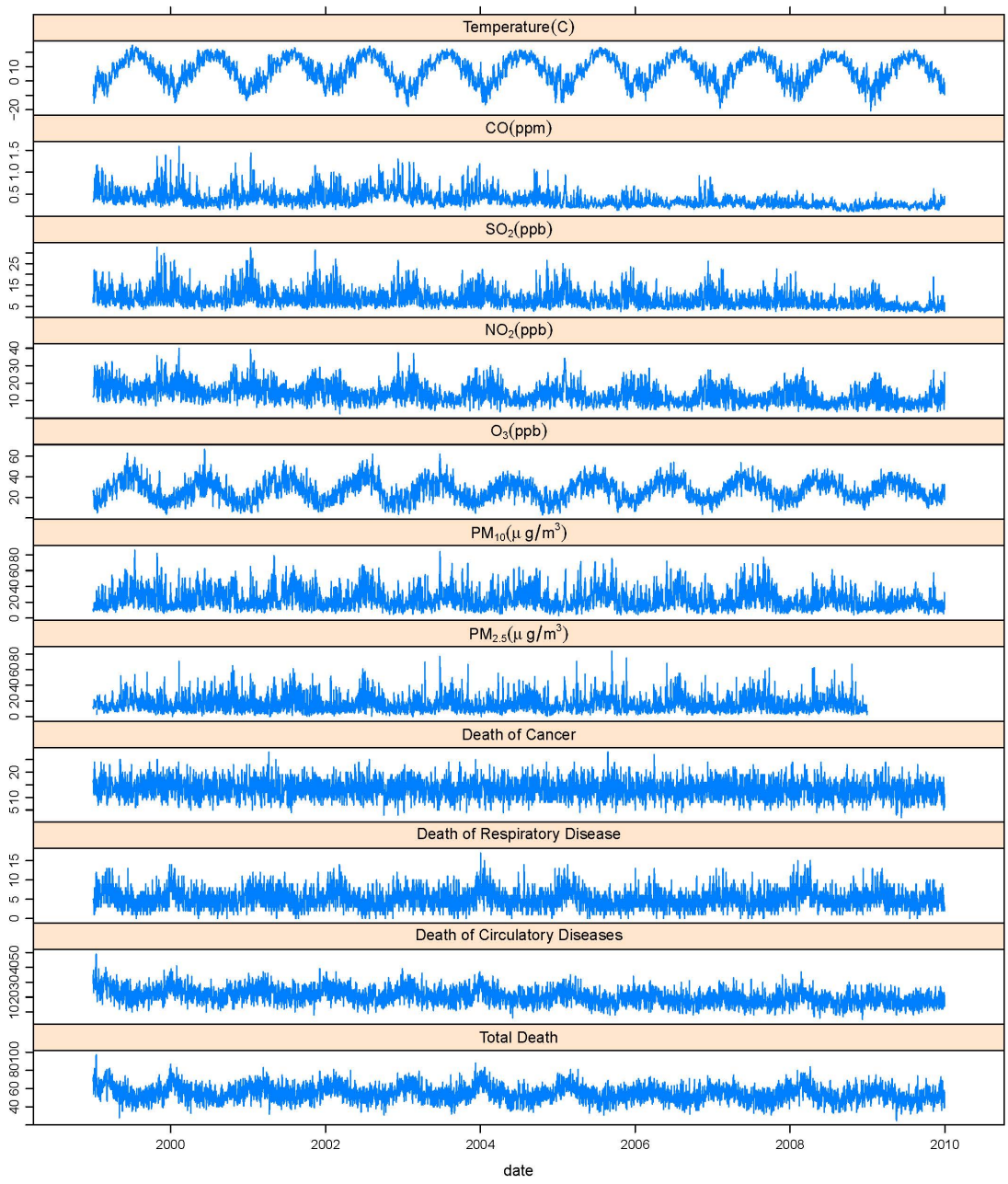
In this section, we are going to show an application of spatiotemporal Kriging in predicting long-term variations of air pollutants in the Pittsburgh region.

Measurements for six air pollutants from 1999 to 2009 were collected from 61 PM_{10}, 47 $PM_{2.5}$, 14 NO_2, 56 SO_2, 39 O_3 and 22 CO routine monitors from multiple networks (e.g. AQS), which were located mainly in western Pennsylvania and eastern Ohio State. Daily temperatures were taken from the Global Historical Climatology Network database provided by NOAA. Air monitors were spatially clustered in the Pittsburgh region (Figure 1), while climate monitors were more evenly distributed. The time series of daily averaged values of all monitors for temperature and six air pollutants are shown in the first six plots in Figure 2. Before Kriging interpolation, some of the air pollutants were transformed to be normally distributed as shown in Table 1.



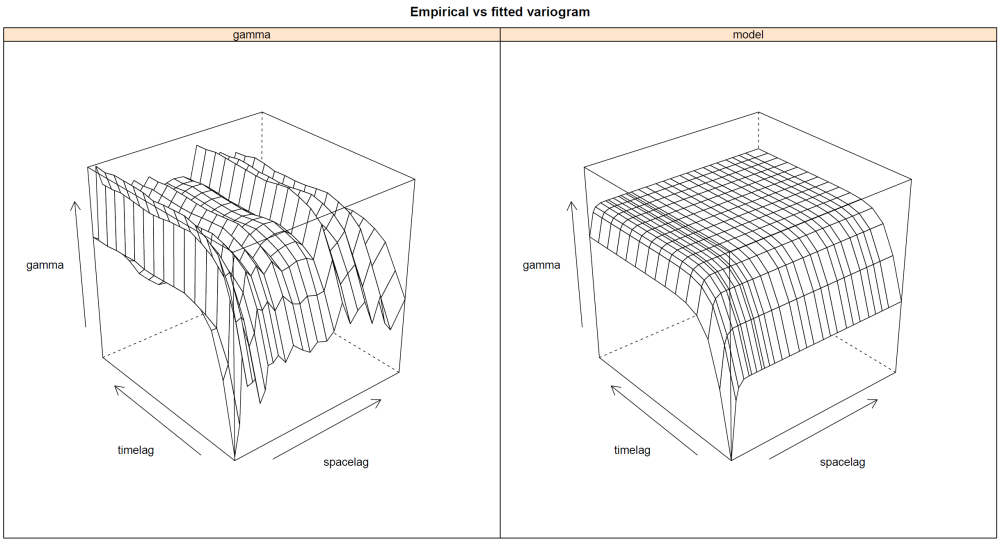
We also show the boundaries of three geographic areas using different colors: city of Pittsburgh (yellow), Allegheny County (blue) and our study domain, which includes the Washington, Allegheny and Westmoreland counties (gray).

Figure 1: Locations of air monitors.

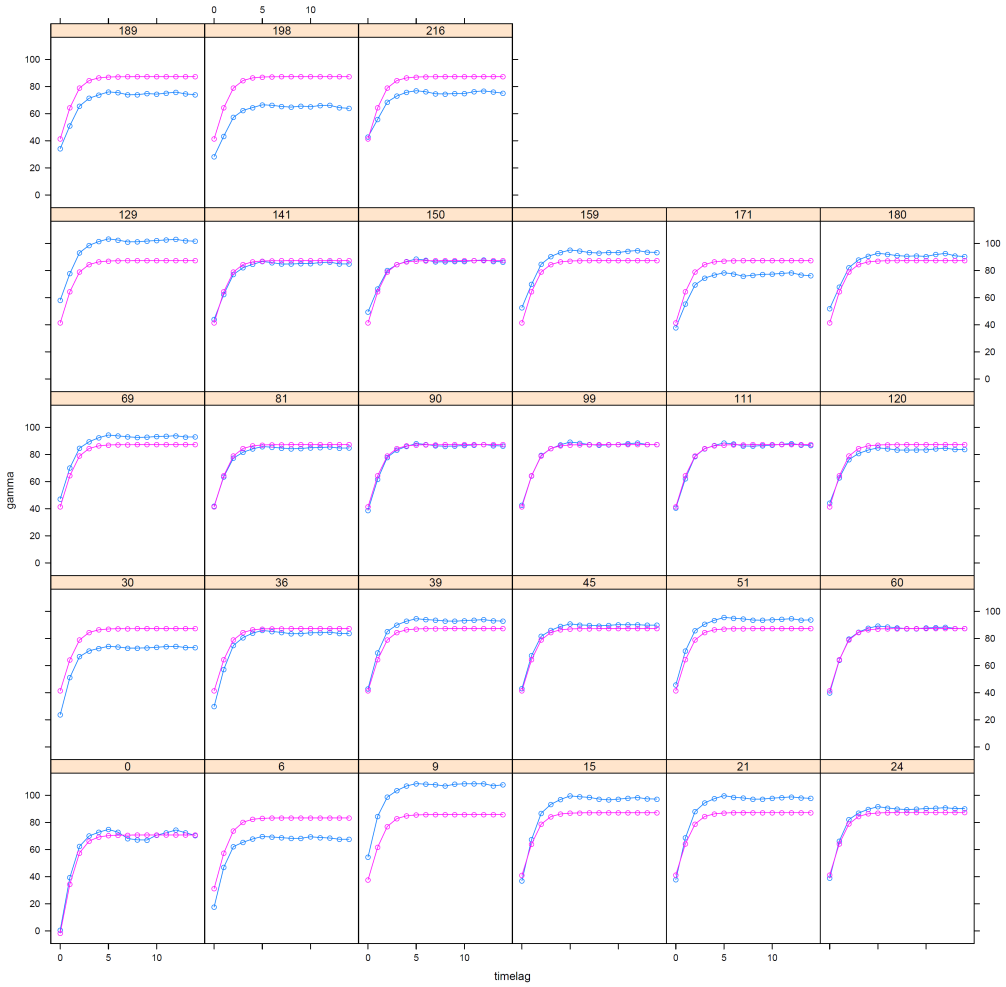


For air pollutants and temperature, the time-series plots display daily averages of all monitors; for mortalities, the time-series plots display daily aggregated counts over Allegheny, Washington and Westmoreland counties.

Figure 2: Time-series plot of air pollutants [for Chapter 2 and 5] and selected cause of death [for Chapter 5].

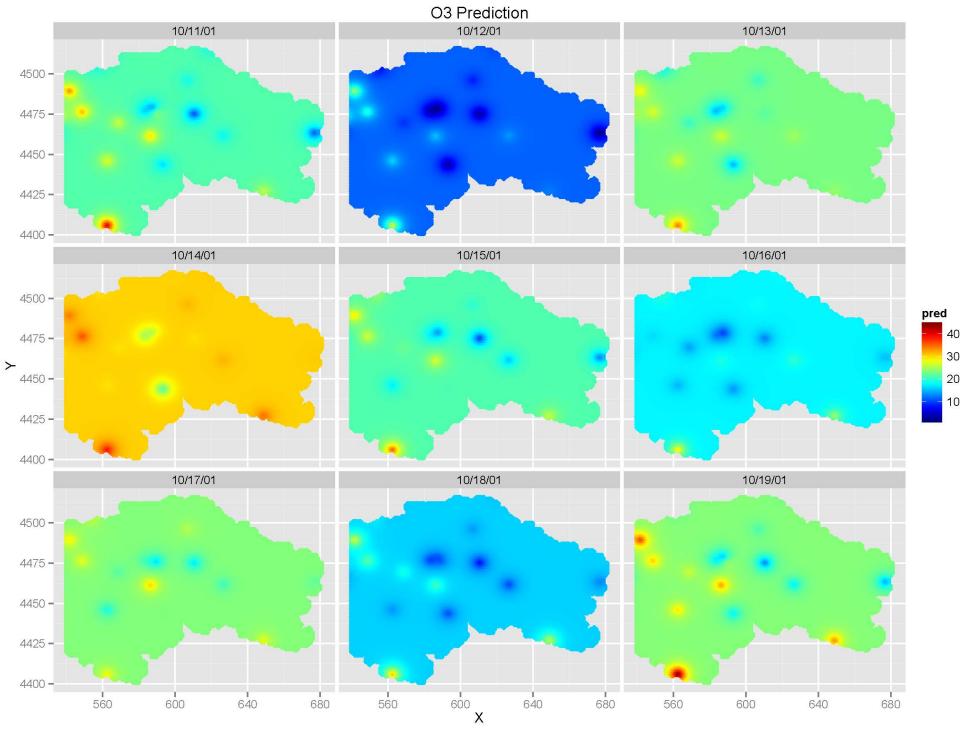


(a) 3D plot: empirical (left) and fitted (right) spatiotemporal variograms.

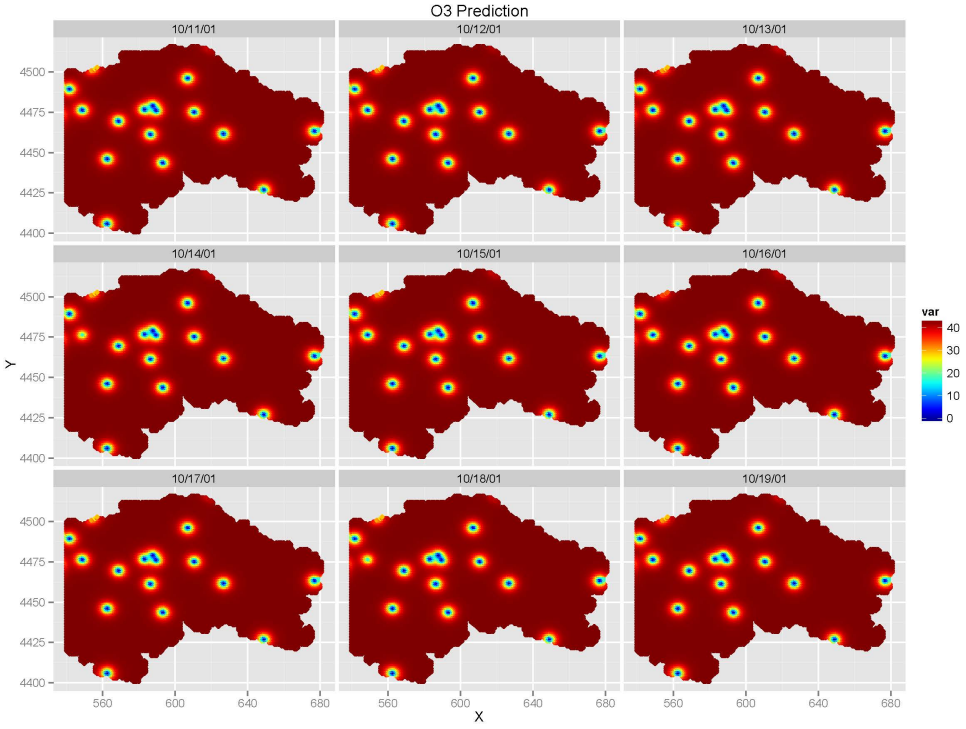


(b) Empirical (blue) and fitted (red) variograms by timelag (day) grouped by specific spacelag (km) values.

Figure 3: Empirical and fitted product-sum variograms for O_3.



(a) Spatiotemproal prediction of O_3



(b) Spatiotemporal Krging variance of O_3

Figure 4: An example of Spatiotemporal Kriging: predictions (a) and their variance (b) for O_3 on $1km \times 1km$ grids for nine consecutive days.

Table 1: 10 fold cross-validation of Spatiotemporal Kriging.

Pollutants	Transform	RMSE$^{\#}$	RMSE*	Variance Explained$^{\#}$	Normalized RMSE$^{\#}$
$PM_{2.5}$	log	0.3051	21.1960	74.77%	50.2
PM_{10}	log	0.2776	7.1547	80.09%	44.6
O_3	identity	3.9826	–	87.88%	34.8
SO_2	log	0.5643	4.4656	64.14%	59.9
NO_2	sqrt	0.3845	2.7869	86.11%	37.3
CO	identity	0.2171	–	60.56%	62.8

Cross-validations are done based on transformed values; $*$ Cross-validations are done based on original scales.

In order to evaluate daily exposure to air pollutants, we applied spatiotemporal Kriging with a product-sum covariance function as shown in Section 2.1.4 to interpolate monitoring data into a regular $1km \times 1km$ grid. The empirical and fitted variograms were calculated using the R package *gstat* [Edzer J. Pebesma, 2004] and an example for O_3 is shown in Figure 3. Based on the graphical interpretation of the variograms, we selected a time window of seven days ($\delta = 7$) for Equation 2.8, as the variograms always reach the sill within seven days as shown in Figure 3(b). For a specific predicted point-value, the spatiotemporal Kriging interpolation combines not only its spatially neighboring monitors but also measurements backward and forward within seven days. An example of nine consecutive days' predictions and their variance for O_3 is shown in Figure 4, and illustrates that Kriging accuracy decreases quickly when leaving a monitoring location. The overall accuracy of spatiotemporal Kriging is evaluated using 10-fold cross-validation as shown in Table 1 and Figure 5. According to the interpolation and cross-validation results, spatiotemporal Kriging is limited in predicting spatiotemporal variations of air pollutants using routine monitoring data because it captures less large-scale spatial variation due to the potential overly smooth Kriging predictions.

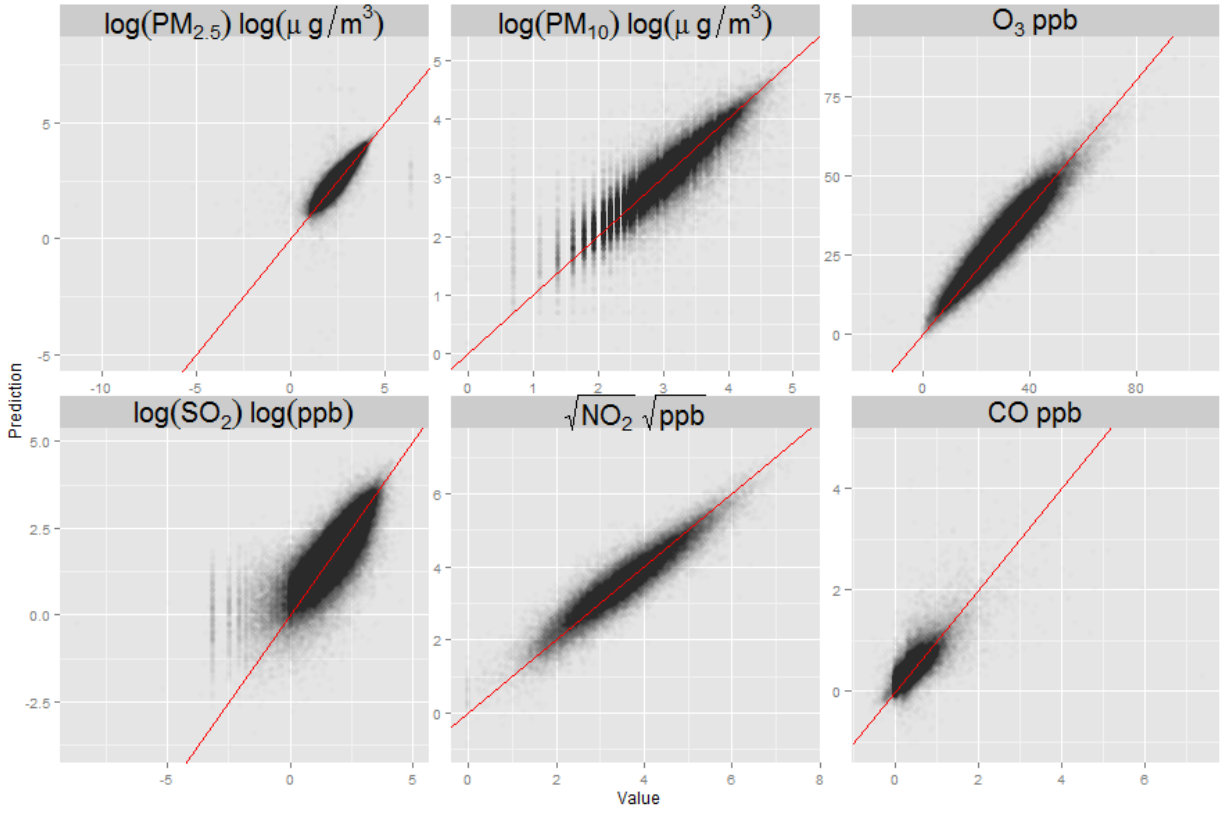


Figure 5: Cross-validation of spatiotemporal Kriging of six air pollutants.

2.2 SPATIAL AND SPATIOTEMPORAL LATTICE PROCESS

Unlike air monitors, most of epidemiological data are collected by areal units such as census blocks, census tracts, USPS ZIP code tabulate areas (ZCTAs), cities, counties and so on. For those scenarios, spatial lattice processes are more reasonable than spatial point processes. (However, researchers can also apply spatial point processes to small areal data, for example, satellite measurements, through using the centroid's coordinate as the location of each areal measurement.) The most popular spatial processes are Conditional Autoregressive (CAR) and Simultaneously Autoregressive (SAR) models. In this section we will briefly review the two lattice process and its extension in modeling spatiotemporal data. In the last section, we will use a study to associate socioeconomic factors with risk of cancer incidence in Allegheny County from 2000 to 2011 as an example to illustrate CAR model.

2.2.1 Spatial Lattice Process and Spatial Hierarchical Model

Spatial processes, CAR and SAR models have been widely applied in spatially modeling normally distributed data in economic, ecological and epidemiological areas. Their modifications, known as hierarchical models, have been developed for count and survival data [Anselin, 1982, Stern and Cressie, 2000, Lichstein et al., 2002, Banerjee et al., 2003, Jin et al., 2005]. First consider a set of spatial measurements $\{x(A_i) : A_i \in (A_1, \cdots, A_n)\}$, where $A_1, \cdots, A_n$ are sub-areas of an irregular lattice area $\boldsymbol{A}$. A spatial process aims to model the dependence structure between those spatial measurements $x(A_i)$. In most scenarios, similar values of spatial data trend to be clustered, which is known as positive spatial autocorrelation and can be tested using statistics such as Moran's I and Geary's C statistics [Anselin, 1995]. In a SAR model, spatial autocorrelation is modeled as a spatial moving average process:

$$x_i = \mu_i + \sum_{j=1}^{n} b_{ij}(x_j - \mu_j) + \epsilon_i,\ \boldsymbol{\epsilon} = [\epsilon_1, \cdots, \epsilon_n]' \sim N_n(\mathbf{0}, \boldsymbol{\Lambda}_s),\ \boldsymbol{\Lambda}_s = \text{diag}(\xi_1^2, \cdots, \xi_n^2)$$

where b_{ij} is a moving average weight and equals 0, if A_i and A_j are not spatial neighbors. The model is referred to as *simultaneous autoregressive* because ϵ_i is correlated with $\{x_j : j \neq i\}$.

Therefore, the joint distribution of a SAR model can be written as a multivariate Gaussian distribution:

$$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, (\boldsymbol{I} - \boldsymbol{C}_s)^{-1}\boldsymbol{\Lambda}_s(\boldsymbol{I} - \boldsymbol{C}_s)^{-T}),\ (\boldsymbol{C}_s)_{ij} = b_{ij}, \tag{2.9}$$

if the inverse of the matrix $(\boldsymbol{I} - \boldsymbol{C}_s)$ exists. Unlike the SAR model, the CAR model assumes $x_i, i = 1, \cdots, n$ are independent of each other conditioning on their neighbors:

$$x_i | x_j \in \{j : j \neq i\} \sim N\left(\mu_i + \sum_{j=1}^{n} c_{ji}(x_j - \mu_j), \tau_i^2\right),$$

where c_{ji} is a spatial weight defined in a similar way as b_{ji}. According to Brooks Lemma, the joint distribution has been developed as a Markov random field [Besag, 1974]:

$$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, (\boldsymbol{I} - \boldsymbol{C}_c)^{-1}\boldsymbol{\Lambda}_c),\ (\boldsymbol{C}_c)_{ij} = c_{ij},\ \boldsymbol{\Lambda}_c = \text{diag}(\tau_1^2, \cdots, \tau_n^2), \tag{2.10}$$

if the inverse of the matrix $(\boldsymbol{I} - \boldsymbol{C}_c)$ exists. The spatial weights matrix ($\boldsymbol{C}_s$ or $\boldsymbol{C}_c$) can be specified based on the neighboring structure of the irregular lattice. Let $\boldsymbol{B}$ denote a binary matrix to identify neighbors based on a specific rule, such as sharing the same edge:

$$(\boldsymbol{B})_{ij} = \begin{cases} 1 & \text{if } A_i \text{ and } A_j \text{ are neighbors} \\ 0 & \text{if } A_i \text{ and } A_j \text{ are not neighbors} \\ 0 & \text{if } i = j \end{cases},$$

and let $\boldsymbol{W}$ denote a row weighted matrix of $\boldsymbol{B}$: $(\boldsymbol{W})_{ij} = (\boldsymbol{B})_{ij} / \sum_{j=1}^{n} (\boldsymbol{B})_{ij}$. The spatial weights matrix $\boldsymbol{C}$ can be defined as $\boldsymbol{W}$ directly or $\rho\boldsymbol{W}$ to guarantee the existence of $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$. The detailed comparisons between the two models have been discussed previously [Cressie, 1993, Banerjee et al., 2004, Wall, 2004] and Cressie has shown that any SAR can be presented by a CAR model through the following equation:

$$(\boldsymbol{I} - \boldsymbol{C}_c)^{-1}\boldsymbol{\Lambda}_c = (\boldsymbol{I} - \boldsymbol{C}_s)^{-1}\boldsymbol{\Lambda}_s(\boldsymbol{I} - \boldsymbol{C}_s)^{-T}.$$

However, due to different forms in model specification, SAR is more appropriate for maximum likelihood inference, while CAR framework can be naturally extended to a Bayesian hierarchical model for non-normally distributed data. In the rest of this section, we will briefly illustrate how to construct a hierarchical model with CAR random effect for spatial count data.

In order to model counts, such as disease counts of a set of areas, researchers usually nested a CAR model as a random effect in a Poisson regression model. Let's $y(A_i) : A_i \in (A_1, \cdots, A_n)$ denotes a count measurement of a spatial lattice and $\boldsymbol{X}$ denotes the covariates matrix, a hierarchical model can be constructed as

$$\begin{aligned} \boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\lambda}),\ \log(\lambda_i) = \theta_i,\ i = 1, \cdots, n; \\ \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\phi},\ \boldsymbol{\phi} \sim \text{CAR}(\cdot|\boldsymbol{\psi}), \end{aligned} \tag{2.11}$$

where the random effect $\boldsymbol{\phi}$ with a conditional autoregressive multivariate normal distribution controls the spatial autocorrelation in count data $\boldsymbol{y}$. The inference of this hierarchical model is usually performed using Bayesian MCMC methods. Both `Winbugs` [Lunn et al., 2000] and some R packages, e.g., `CARBayes` [Lee, 2013] provide mature MCMC simulation for such hierarchical models.

2.2.2 Spatiotemporal Hierarchical Model

For spatiotemporal lattice data, such as a set of temporally repeated measurements of a spatial lattice, the temporal autocorrelation can be controlled simultaneously as spatial autocorrelation through simply adding another random effect of a one-dimensional CAR, autoregressive (AR) or moving average (MA) multivariate distribution into a hierarchical model (e.g. Equation 2.11):

$$\begin{aligned} (y_{11}, \cdots, y_{n1}, \cdots\cdots, y_{1m}, \cdots, y_{nm})' \equiv \boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\lambda}),\ \log(\lambda_{ij}) = \theta_{ij}; \\ \theta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \phi_i + \varphi_j, \\ (\phi_j, \cdots, \phi_m)' \equiv \boldsymbol{\phi} \sim \text{CAR}(\cdot|\boldsymbol{\psi}), \\ (\varphi_j, \cdots, \varphi_m)' \equiv \boldsymbol{\varphi} \sim \text{AR}(\cdot|\boldsymbol{\nu}) \text{ or } \text{MA}(\cdot|\boldsymbol{\nu}) \text{ or } \text{CAR}(\cdot|\boldsymbol{\nu}); \\ i = 1, \cdots, n,\ \ j = 1, \cdots, m; \end{aligned} \tag{2.12}$$

where i and j denote the spatial and temporal indexes of spatiotemporal measurements $\boldsymbol{y}$, $\boldsymbol{x}_{ij}$ denotes the covariates for y_{ij}, $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ denote the spatial and temporal random effects and $\boldsymbol{\psi}$ and $\boldsymbol{\nu}$ denote the tuning parameters in spatial and temporal random effects. Researchers have constructed complicated spatiotemporal hierarchical models through modifications of the above Equation 2.12. For example, the tuning parameter for spatial autocorrelation $\boldsymbol{\psi}$

can be defined as time-varying to allow spatial autocorrelation to change with time. More complex spatiotemporal hierarchical models have been developed using more complicated covariance structures, for example, models that include the interaction between temporal and spatial autocorrelations [Waller et al., 1997, Mariella and Tarantino, 2010].

2.2.3 An Example of a Conditional Autoregressive Model: Association between Socioeconomic Factors and Risk of Cancer in Allegheny County, 2000-2011

In this section, we illustrate the CAR model by applying it to associate socioeconomic factors to census tract level cancer incidence in Allegheny County, Pennsylvania. According to cancer statistics published by the Pennsylvania Department of Health (https://apps.health.pa.gov/EpiQMS/asp/ChooseDataset.asp)[1], we list the five top cancers in Allegheny County and the state of Pennsylvania in 2000 and 2011 (Table 2) and plot the yearly time series of the standard age-adjusted rate of cancer incidences as Figure 6. In the following study, we are going to explore the socioeconomic factors' effects on the five top cancers in Allegheny County.

Individual cancer registry records, including sex, age, time at diagnosis, race, marital status, birth date and place, address at diagnosis time and IDC-10 codes for cancer site, behavior and histology from 2000 to 2011 were retrieved from the Pennsylvania Department of Health. The specific type of invasive cancer was identified by primary site, behavior and histology code using SEER's rocode protocol (http://seer.cancer.gov/analysis/). Each cancer record was assigned to the 2000 census tracts for Allegheny County according to location coordinates geocoded by the address information. Census tract level demographic data by sex, age and race groups and socioeconomic data including mean age, median household income, median family income, percent of family in poverty, number of individuals in poverty, percent of individuals in poverty, percent of unemployed males, percent with less than high school education, percent of female headed household, and percent of public assistance were also collected from 2000 census (http://www.census.gov/). As the socioeconomic factors are cor-

[1]These data were provided by the Bureau of Health Statistics and Research, Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Table 2: Top five cancer types in Allegheny County and the state of Pennsylvania for 2000 and 2011.

Rank	2000				2011			
	Pennsylvania (%)		Allegheny (%)		Pennsylvania (%)		Allegheny (%)	
1	PROSTATE	15	BRONCHUS & LUNG	15.2	BRONCHUS & LUNG	13.5	BRONCHUS & LUNG	15.3
2	BRONCHUS & LUNG	14.3	FEMALE BREAST	14.6	FEMALE BREAST	13.5	FEMALE BREAST	14.1
3	FEMALE BREAST	14.2	PROSTATE	14.2	PROSTATE	13.1	PROSTATE	11
4	COLON & RECTUM	12.7	COLON & RECTUM	12.6	COLON & RECTUM	9	COLON & RECTUM	8.6
5	URINARY BLADDER	5.2	URINARY BLADDER	5.5	URINARY BLADDER	5.2	URINARY BLADDER	5.4

related with each other, to simplify, we summarized six representative socioeconomic factors and created a Socioeconomic Status (SES) index , which have been associated with various types of cancer in previous studies [Yost et al., 2001, Robert et al., 2004, Cheng et al., 2009]:

$$\begin{aligned} Z = \; & Z_{\text{median household income}} - Z_{\text{percent with less than high school education}} - Z_{\text{percent of unemployed males}} \\ & - Z_{\text{percent in poverty}} - Z_{\text{percent of public assistance}} - Z_{\text{percent of female headed household}} \end{aligned}$$

$$\text{SES index} = \frac{Z - \min(Z)}{\max(Z) - \min(Z)} \times 100\%,$$

where Z denotes the measurements of these socioeconomic factors. We also collected neighborhood levels of smoking and obesity indicators from the Behavioral Risk Factor Surveillance System (BRFSS) for 2000 and assigned them to the appropriate census tract.

In order to compare cancer risk between various census tracts and adjust potential confounding effects of age, sex and race, we calculated adjusted standardized incidence ratio (SIR). Let $y_{i,k}$ and $p_{i,k}$ denote aggregated cancer counts and total population in i^{th} census tract for the k^{th} demographic group categorized by sex, age and race. The adjusted SIR for i^{th} census tract can then be calculated as:

$$\text{SIR}_i = \frac{y_i}{e_i}, \quad y_i = \sum_k y_{i,k}, \quad e_i = \sum_k p_{i,k} \left(\frac{\sum_i y_{i,k}}{\sum_i p_{i,k}} \right), \tag{2.13}$$

where y_i is the observed count of cancer in i^{th} census tract and e_i is known as the expected count for cancer in the i^{th} census tract and can be used as an offset to adjust age, sex and race in Poisson regression. If the SIR equals one, we can conclude that the cancer risk for

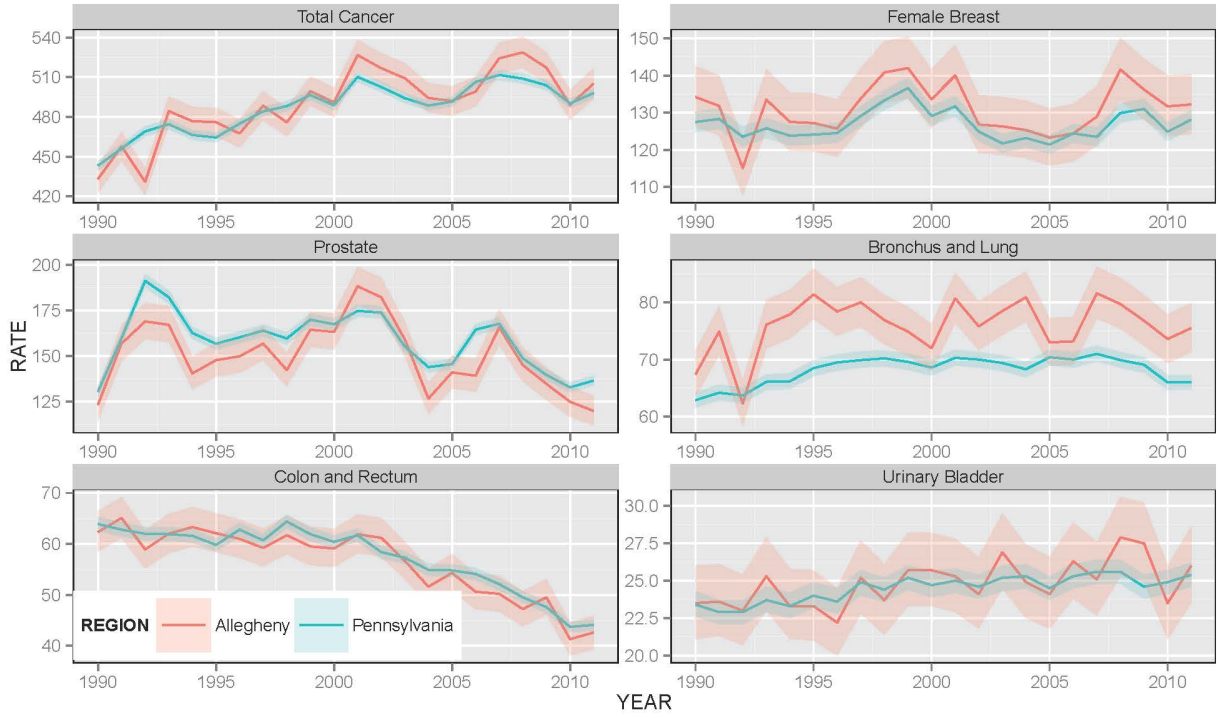


Figure 6: Age-adjusted Rate per 100,000 of total and selected cancers for Allegheny County and the state of Pennsylvania in 2000-2011.

the census tract is equal to the average level in the whole study domain. The census tract level SIRs for the total cancer and five top cancers are shown in Figure 7 and SES index, overweight, percent of obesity and percent of smoking are shown in Figure 8.

According to Section 2.2.1, a spatial hierarchical model for cancer incidence can be constructed as follows:

$$y_i \sim \text{Poisson}(\lambda_i e_i), \quad \theta_i = \log(\lambda_i), \quad [\theta_1, \cdots, \theta_n]' = \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\phi}, \quad \boldsymbol{\phi} \sim \text{CAR}(\cdot|\boldsymbol{\psi}). \quad (2.14)$$

In the model, the relative risk (RR) for cancer types can be interpreted as $\exp(\beta)$. The model inference performed using MCMC from R package `CARBayes` [Lee, 2013]. We first associated each socioeconomic factor separately with top five cancer SIRs (Table 7) and then selected the SES index, percent of obesity, percent of overweight and percent of smoking from all of the covariates as covariates in the final model to estimate top five cancer risks.

In Table 4, we also compare the spatial CAR model with the independent Poisson regression using the deviance information criterion (DIC). We find that in most scenarios, the CAR models are better than Poisson regressions (as the DICs are smaller for CAR models), except for colon and rectum cancer. Finally we compared the fitted and observed SIRs and also identified the statistically significantly higher census tracts, whose 95% quantile for the posterior distribution the fitted cancer counts were higher than the expected counts in Figure 9.

The study only aims to illustrate the application of spatial CAR model and its performance in epidemiology. We cannot over-interpret the modeling results in this study, as it is limited in several respects. Firstly, all the socioeconomic factors were collected in 2000 but may not represent the exposed risk factors in cancer developments due to a 20 to 30 year period of latency for various types of cancer. Secondly, a more representative SES index can be developed using methods such as principle component analysis rather than the simple linear combination used in this study. Thirdly, smoking and obesity risk factors are collected in neighborhoods, which are misaligned with census tracts and thus may cause potential exposure misclassification.

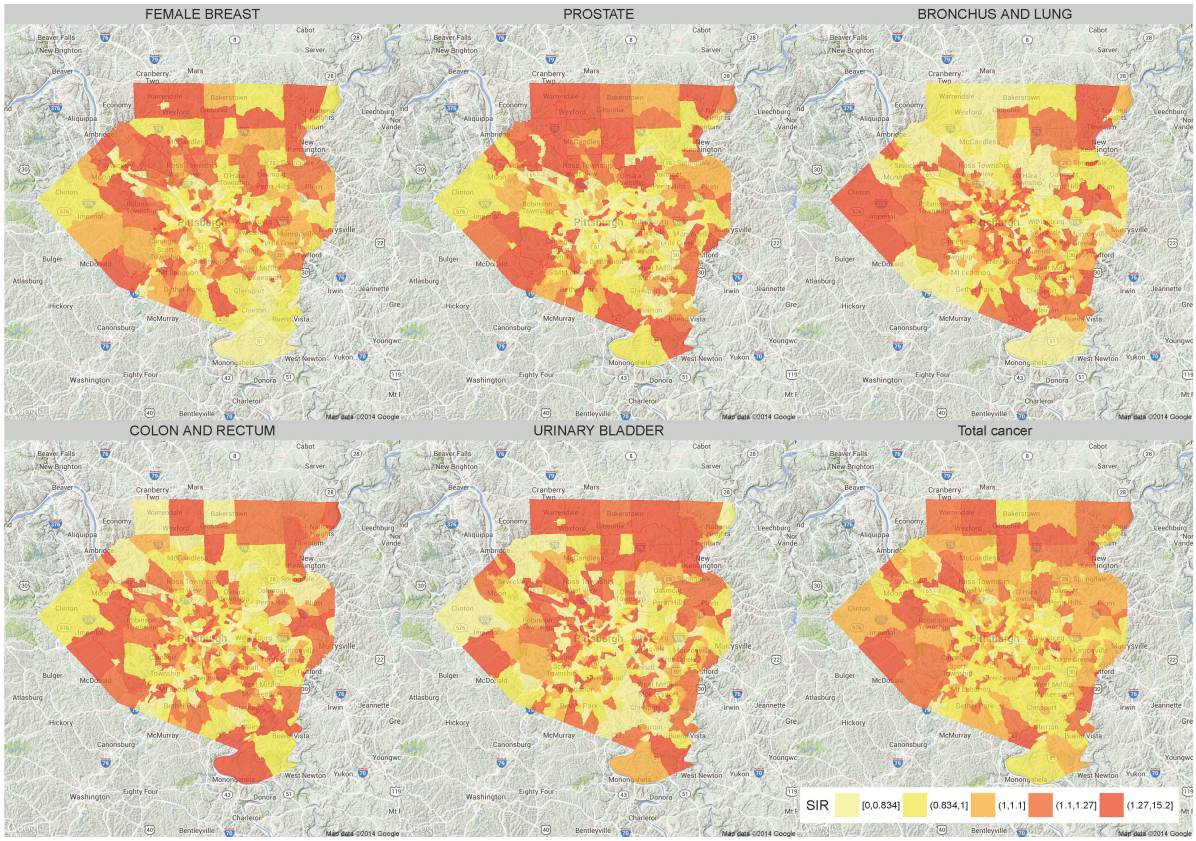
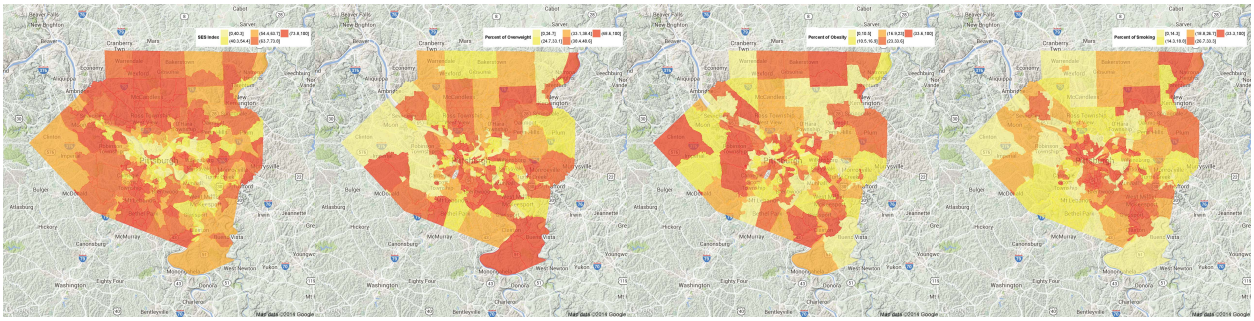


Figure 7: Age, sex and race adjusted SIRs of total cancer and five top cancers in census tracts of Allegheny County.



(a) SES index (b) Percent of overweight (c) Percent of obesity (d) Percent of smoking

Figure 8: SES index, obesity and smoking in census tracts of Allegheny County.

Table 3: Relative risks for the single CAR model between socioeconomic factors and cancer SIRs.

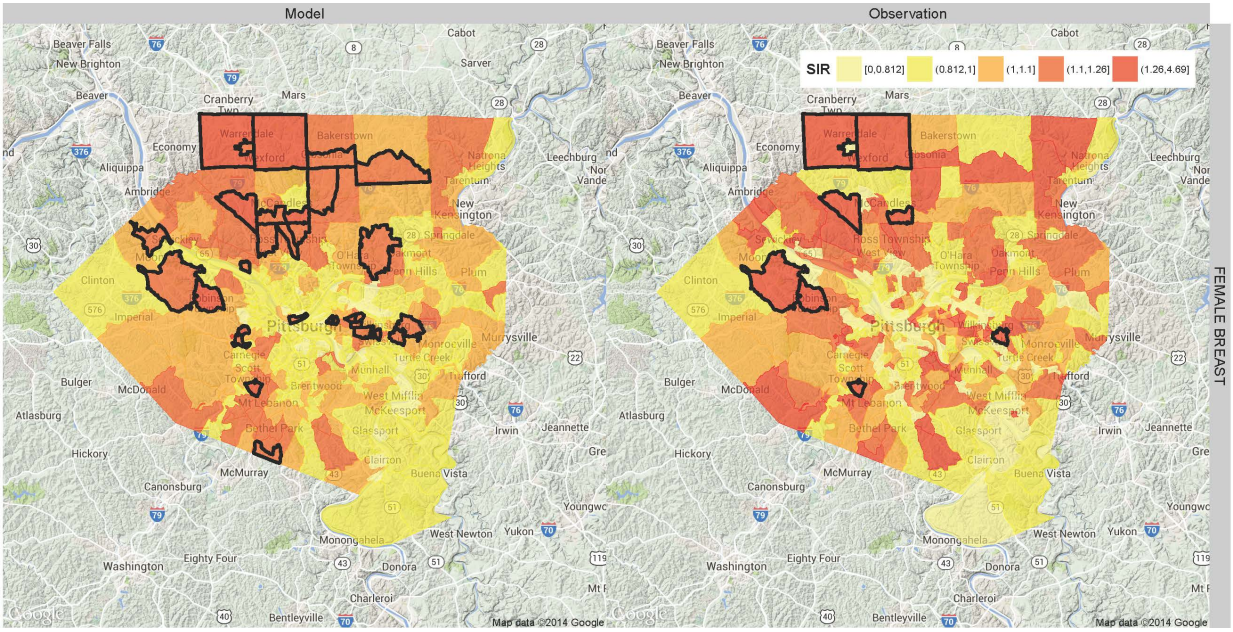
Lifestyle factors	RR (per unit change)				
	Breast	Prostate	Lung	Colon	Bladder
Percent of overweight	1.0011	1.0012	1.0004	1.0015	1.0009
Percent of obesity	0.9982	0.9999	1.0017	0.9992	1.0002
Percent of overweight and obesity	0.9997	1.0011	1.0016	1.0009	1.0010
Percent of smoking	0.9964	0.9954	1.0022	0.9993	0.9982
Median household income	1.0060	1.0068	0.9950	0.9998	1.0016
Median family income	1.0048	1.0050	0.9956	0.9990	1.0006
Percent of family in poverty	0.9908	0.9914	1.0031	0.9962	0.9915
Number of individuals in poverty	0.9997	0.9997	1.0001	0.9998	0.9998
Percent of individuals in poverty	0.9910	0.9904	1.0027	0.9948	0.9926
Percent of public assistance	0.9760	0.9767	1.0015	0.9890	0.9841
Percent of unemployed males	0.9914	0.9949	1.0011	0.9928	0.9934
Percent with less than high school	0.9850	0.9828	1.0080	0.9982	0.9943
Percent of female headed household	0.9873	0.9884	1.0026	0.9989	0.9951
SES Index	1.0072	1.0072	0.9968	1.0011	1.0019

Significant positive RRs are highlighted in red, while significant negative are blue.

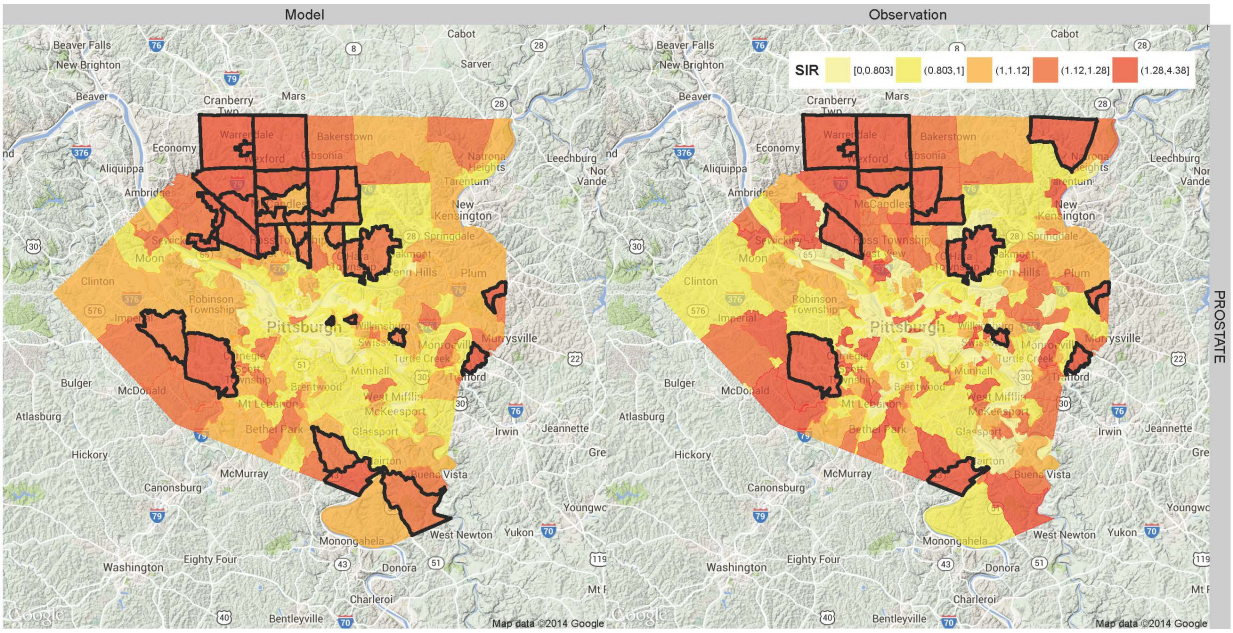
Table 4: Associating cancer SIRs with the SES index, percent of overweight, percent of obesity and percent of smoking using the spatial CAR model vs the independent Poisson model.

y	x	Independent Poisson					CAR Model				
		$\hat{\beta}$	2.5%CI	97.5%CI	RR	DIC	$\hat{\beta}$	2.5%CI	97.5%CI	RR	DIC
FEMALE	SES Index	0.0069	0.0056	0.0083	1.0069	2650.60	0.0070	0.0056	0.0085	1.0070	2641.36
BREAST	Percent of overweight	0.0005	-0.0009	0.0019	1.0005		0.0008	-0.0006	0.0023	1.0008	
	Percent of obesity	0.0002	-0.0014	0.0019	1.0002		0.0003	-0.0015	0.0019	1.0003	
	Percent of smoking	-0.0009	-0.0028	0.0010	0.9991		-0.0007	-0.0027	0.0012	0.9993	
PROSTATE	SES Index	0.0075	0.0060	0.0091	1.0075	2665.61	0.0072	0.0054	0.0089	1.0072	2658.61
	Percent of overweight	0.0017	0.0001	0.0033	1.0017		0.0019	0.0003	0.0036	1.0019	
	Percent of obesity	0.0023	0.0004	0.0042	1.0023		0.0024	0.0006	0.0042	1.0024	
	Percent of smoking	-0.0024	-0.0045	-0.0003	0.9976		-0.0023	-0.0043	-0.0002	0.9977	
BRONCHUS	SES Index	-0.0029	-0.0045	-0.0013	0.9971	2837.22	-0.0025	-0.0043	-0.0007	0.9975	2829.63
AND	Percent of overweight	0.0007	-0.0011	0.0025	1.0007		0.0009	-0.0008	0.0027	1.0009	
LUNG	Percent of obesity	0.0009	-0.0010	0.0030	1.0009		0.0015	-0.0005	0.0035	1.0015	
	Percent of smoking	0.0018	-0.0004	0.0039	1.0018		0.0013	-0.0009	0.0035	1.0013	
COLON	SES Index	0.0008	-0.0008	0.0024	1.0008	2567.42	0.0011	-0.0007	0.0029	1.0011	2565.22
AND	Percent of overweight	0.0018	0.0002	0.0036	1.0018		0.0016	-0.0001	0.0034	1.0016	
RECTUM	Percent of obesity	0.0004	-0.0015	0.0024	1.0004		0.0003	-0.0017	0.0024	1.0003	
	Percent of smoking	-0.0000	-0.0023	0.0022	1.0000		-0.0002	-0.0024	0.0020	0.9998	
URINARY	SES Index	0.0020	0.0000	0.0040	1.0020	2172.73	0.0020	-0.0000	0.0040	1.0020	2177.93
BLADDER	Percent of overweight	0.0014	-0.0006	0.0035	1.0014		0.0014	-0.0006	0.0034	1.0014	
	Percent of obesity	0.0015	-0.0009	0.0040	1.0015		0.0015	-0.0009	0.0039	1.0015	
	Percent of smoking	-0.0008	-0.0036	0.0020	0.9992		-0.0008	-0.0035	0.0019	0.9992	

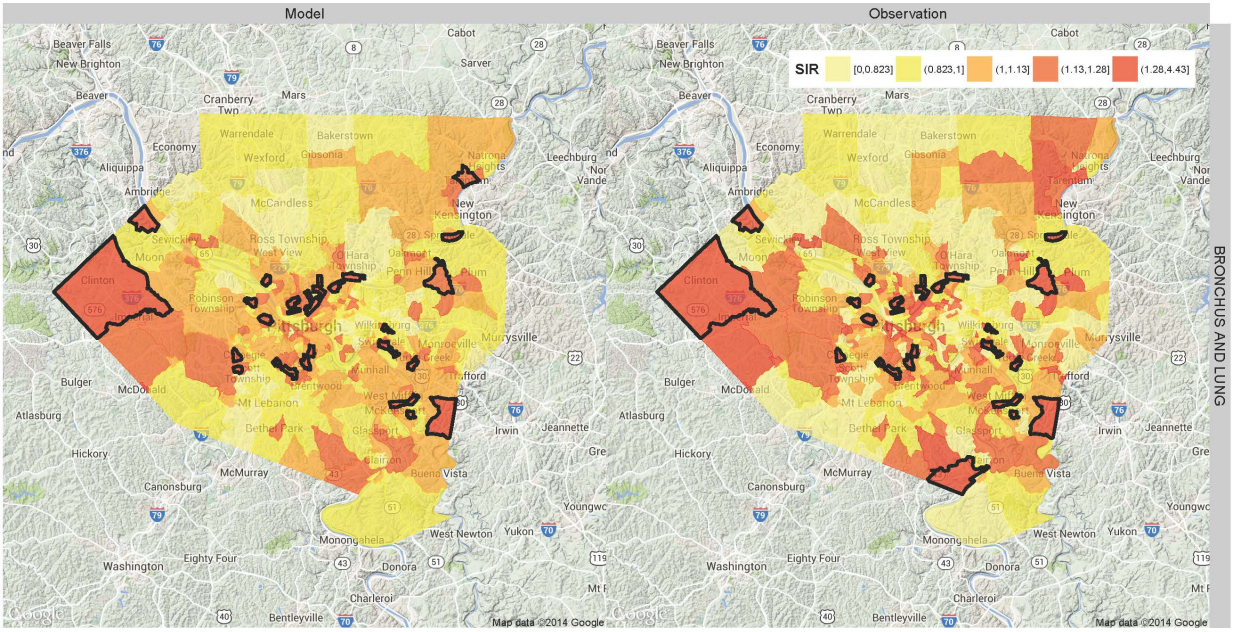
Statistically significant positive regression coefficients are highlighted in red, while the negative ones are blue.



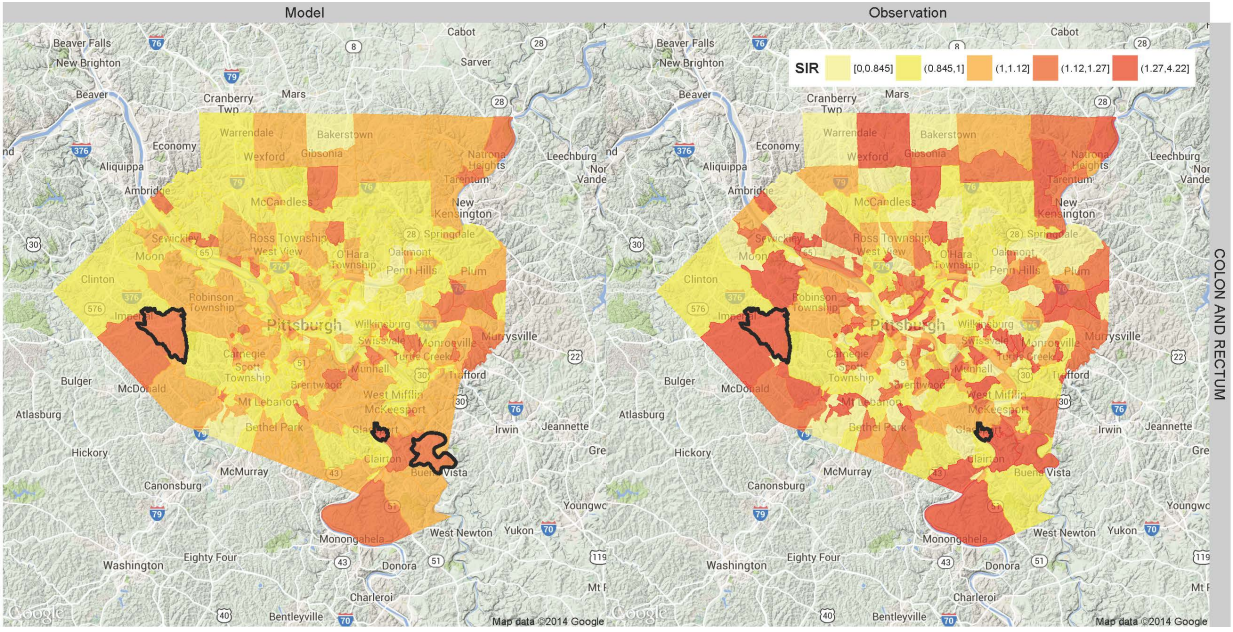
(a) Breast



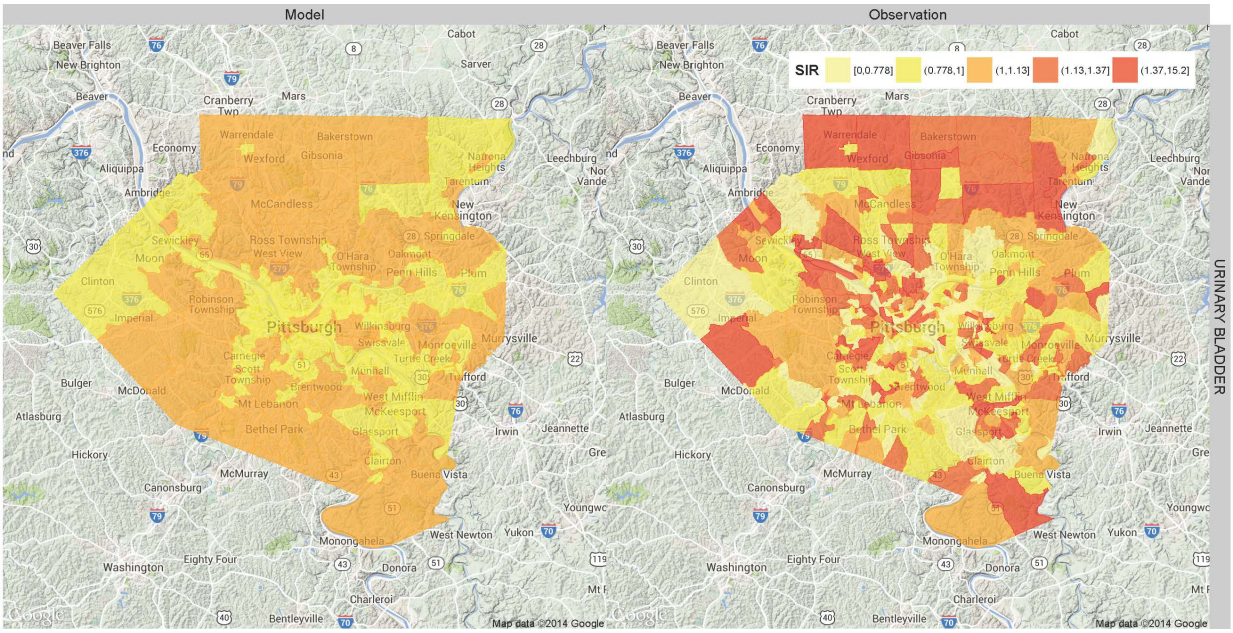
(b) Prostate



(c) Bronchus & Lung



(d) Colon & Rectum



(e) Bladder

Compare model fitted SIR (left column) with observed SIR (right column) for top five types of cancer in Allegheny County. The significant higher census tracts are marked by black polygons.

Figure 9: Observed SIRs vs Predicted SIRs by CAR models.

3.0 ESTIMATING SPATIOTEMPORAL VARIATIONS OF $PM_{2.5}$ MASS CONCENTRATION USING MONITORING SATELLITE AEROSOL OPTICAL DEPTH DATA OVER THE PITTSBURGH REGION, 2001-2008

In Section 2.1.4, we have found out that spatiotemporal Kriging is not able to capture large scale spatial variations of $PM_{2.5}$ mass concentration and its accuracy is relative lower among all types of pollutants. Therefore, in this chapter, we will develop a study to generate better spatiotemporal predictors for $PM_{2.5}$ through using satellite measurements of aerosol optical depth (AOD).

3.1 INTRODUCTION AND DATA

3.1.1 Introduction

Many epidemiological studies have associated short-term exposure to particulate matter with aerodynamic diameter $\leq 2.5\mu m$ ($PM_{2.5}$) with adverse health outcomes, including cardiovascular and respiratory diseases [Dominici et al., 2006, Glad et al., 2012, Peng et al., 2009]. Assessing exposure at the individual-level is essential for accurately evaluating the health risks of $PM_{2.5}$ [Gamble, 1998]. However, because of the limited number of routine monitoring stations, many previous time-series studies have assigned the same exposure level for a group of residents [Pope III et al., 1995], thus ignoring the spatial variation of air pollutants. As the chemical components of, and sources contributing to $PM_{2.5}$ could be highly heterogeneously distributed in a city [Kim et al., 2005], and within-city health effects of $PM_{2.5}$ were possibly larger than between-city effects [Jerrett et al., 2005, Miller et al., 2007],

ignoring spatial variation could lead to potential exposure misclassification in time-series epidemiological studies [Pinto et al., 2004].

Previous studies applied the land use regression (LUR) model [Clougherty et al., 2008, Henderson et al., 2007] and spatial interpolation methods (e.g. Kriging and its extensions [Jerrett et al., 2005]) to estimating spatial variations of $PM_{2.5}$. LUR typically uses traffic indicators, emission sources, land cover types to associate with station monitoring data and has performed better with traffic-related air pollutants (e.g. NO_2) than $PM_{2.5}$ [Henderson et al., 2007]. Applications of LUR have usually ignored spatial auto-correlation in contrast to the use of universal Kriging [Mercer et al., 2011]. Both ordinary Kriging and universal Kriging have been used to interpolate $PM_{2.5}$ [Jerrett et al., 2005] and produced best linear unbiased predictors (BLUP) at given locations based on their neighboring monitors. As ordinary Kriging is a univariate process for $PM_{2.5}$ measurements, its spatial resolution is limited by sparsely distributed monitoring stations, while universal Kriging (also known as Kriging with external drift or regression Kriging [Hengl et al., 2004]) allows the inclusion of additional linear predictors such as land use variables.

However, both LUR models and Kriging have methods usually characterized the long-term spatial variations of air pollutants [Beelen et al., 2009, Henderson et al., 2007] but not the temporal variations because of little or no temporal variation in land use variables (e.g. distances to major road) [Hoek et al., 2008] and the potentially complicated spatiotemporal covariance structure of the air pollutants [Bruno et al., 2009]. Satellite AOD, which reflects the vertical column abundance of particulate matter from the earth's surface to the atmospheric top, has been widely used to estimate spatiotemporal patterns of $PM_{2.5}$ [Liu et al., 2009, Paciorek et al., 2008] because of its global spatial coverage and daily temporal resolution. Additionally, as we discussed in Section 2.1.1, a complex spatiotemporal covariance structure can be simplified by assuming (1) a stationarity spatiotemporal process, and (2) a product-sum covariance function.

Both statisticians and environmental scientists have studied the spatiotemporal variations of particulate matter [Choi et al., 2009, De Iaco et al., 2002a, Kumar et al., 2007, Liu et al., 2007, Liu et al., 2009, Paciorek et al., 2008, Sahu et al., 2006]. However, statisticians have focused on modeling the complex spatiotemporal field of the univariate

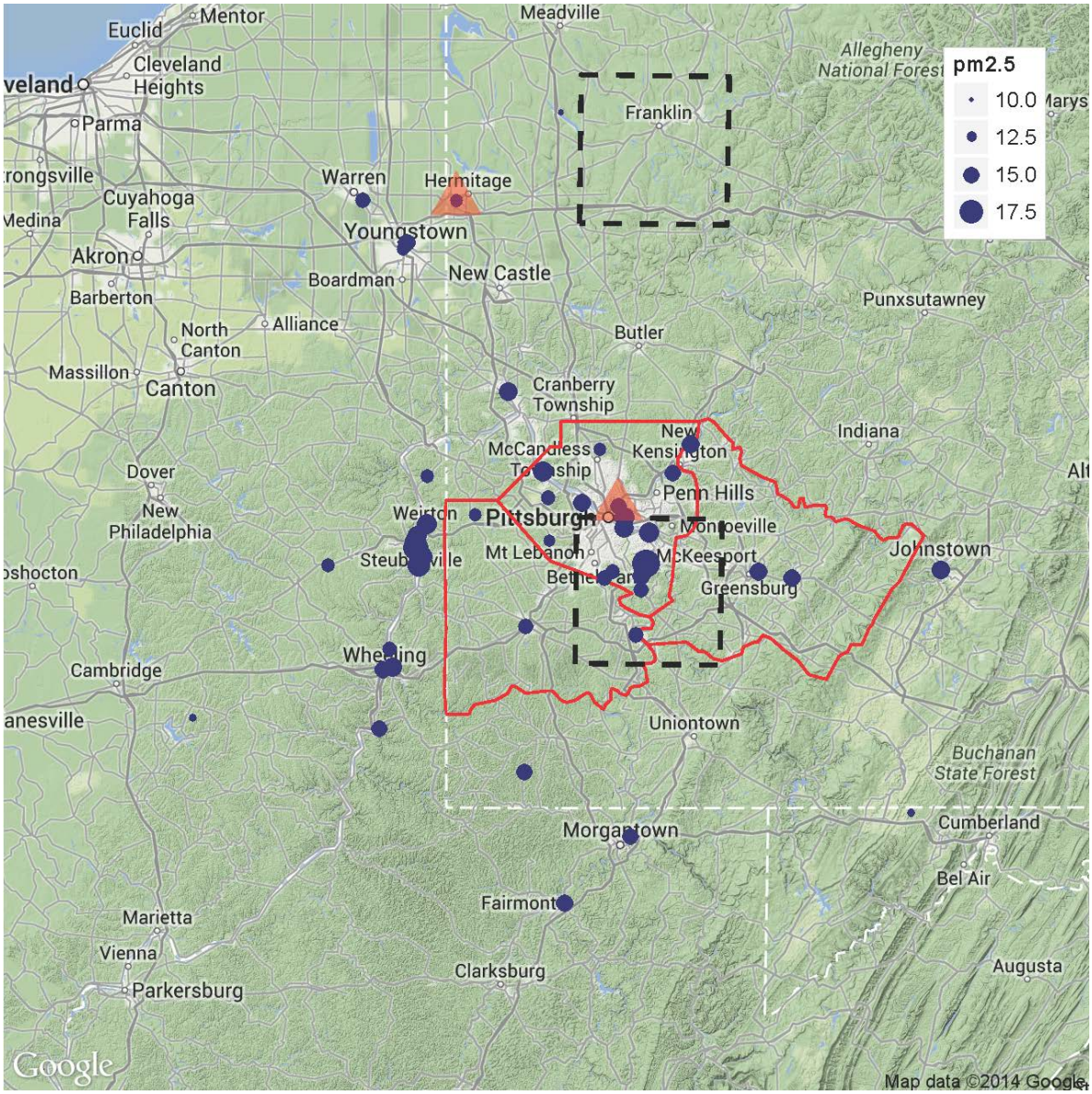
process of particulate matter using hierarchical models, but usually ignored useful predictors such as AOD [Choi et al., 2009, De Iaco et al., 2002a, Sahu et al., 2006]; environmental scientists have applied sophisticated statistical methods such as the generalized additive model (GAM) [Hastie and Tibshirani,] to model the complicated non-linear association between $PM_{2.5}$ and AOD but put less effort in modeling the stochastic components [Liu et al., 2009]. In addition, despite researchers having reported seasonal variation of correlation between $PM_{2.5}$ and AOD (lower in winter, higher in summer and fall) [Paciorek et al., 2008, Zhang et al., 2009], previous statistical analyses have rarely considered a time-varying relationship between AOD and $PM_{2.5}$ in their regression model.

In this section, we will improve the performance of spatiotemporal Kriging in predicting $PM_{2.5}$ by including AOD measurements from the Moderate-resolution Imaging Spectroradiometer (MODIS).

3.1.2 Data Description

We use the same monitoring data for $PM_{2.5}$ mass concentrations as those in Section 2.1.4, whose geographic information are also shown in Figure 10. All measurements were aggregated into a daily average with an exclusion criteria of $> \%10$ missing values. Systematic errors between various measuring methods of $PM_{2.5}$ have been estimated and all non-Federal Reference Method (FRM) samples haven been calibrated to FRM equivalent values using the methods described in our previous paper [Bilonick et al., 2015]. We excluded 39 calibrated values $> 600\mu g/m^3$ as potential outliers. As MODIS launched on NASA satellites from September 18, 1999, in this section, we used 61,346 calibrated $PM_{2.5}$ measurements (2,887 days; an average of 21.25 measurements per day) from January 11, 2001 to December 31, 2009.

The Moderate-resolution Imaging Spectroradiometer (MODIS), operated by NASA, had been launched on two Earth Observing System satellites: Terra (from 1999) and Aqua (from 2002) in earth orbit and was designed to provide information about terrestrial, oceanic, and atmospheric conditions with 36 spectral channels from 0.4 μm to 14.4 μm. AOD measures light extinction integrated over a path which usually means a vertical column from the earth's



The red outlines indicate the study domain (from left to right: Washington, Allegheny and Westmoreland counties); the blue dots display locations of samplers and the size of the dot denotes the average concentration of $PM_{2.5}$ ($\mu g/m^3$) from 1999 to 2011; two dashed squares display two areas used for AOD cross validations ($CV_S^{(center)}$ and $CV_S^{(edge)}$) and two red triangles denotes two monitoring sites for $PM_{2.5}$ cross validations ($CV_S^{(center)}$ and $CV_S^{(edge)}$).

Figure 10: Study domain and locations of routine samplers of $PM_{2.5}$.

surface to the top of the atmosphere by satellite remote sensing. The MODIS team use the ratio of scattering in the red (0.66 μm) and blue (0.47 μm) wavelengths and could choose one of four models (dust aerosol, biomass burning, industrial/urban aerosol, or continental aerosol) according to the geographical location and season to calculate land AOD under the condition of a clear sky without clouds. MODIS AOD has been reported as better correlated with ground-based $PM_{2.5}$ in the eastern and Midwest portion of the US than the rest of the continental United States [Engel Cox et al., 2004].

We obtained the MODIS AOD observations (Level 2; Collection 51; $MOD04\_L2$) which ranged between -0.5 μ and 0.55 μ from the Atmosphere Archive and Distribution System (LAADS, http: //ladsweb.nascom.nasa.gov) with a spatial resolution of $10 \times 10km$ and temporal resolution of 1 day in Pittsburgh. Terra scans the study domain from 3 p.m. to 5 p.m. each day. In total, we collected 309,919 measurements of AOD for 1,955 days (an average of 158 measurements per day) from January 11, 2001 to December 31, 2009.

3.2 STATISTICAL MODEL: TWO-STEP ADDITIVE MIXED EFFECTS MODEL

In this section, we will construct a two-step model to associate AOD with $PM_{2.5}$ mass concentration and predict spatiotemporal variations of $PM_{2.5}$ at a spatial resolution of $1km \times 1km$. In the first stage of the model, we apply spatiotemporal Kriging to AOD data to adjust the spatial misalignment between AOD and $PM_{2.5}$ routine monitors; in the second stage, we construct a varying-coefficient mixed effect model to associate adjusted AOD from the first stage with monitoring $PM_{2.5}$ and to simultaneously control stochastic spatiotemporal random effect in $PM_{2.5}$.

Let $\boldsymbol{z}$ and $\boldsymbol{y}$ denote AOD and $PM_{2.5}$ measurements, respectively. Let ($\boldsymbol{s}_z$, $\boldsymbol{t}_z$) and ($\boldsymbol{s}_y$, $\boldsymbol{t}_y$) denote the spatial and temporal coordinates for AOD and $PM_{2.5}$, respectively. Let ($\boldsymbol{s}_p$,$\boldsymbol{t}_p$) denote spatial and temporal coordinates for a regular spatiotemporal grid. Therefore, in stage 1, AOD is modeled as a multivariate normal distribution with product-sum covariance

function:

$$\boldsymbol{z} \sim N_n(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z), \quad \boldsymbol{\mu_z} = f(\boldsymbol{s}_z, \boldsymbol{t}_z), \quad (\boldsymbol{\Sigma}_z)_{i,j} = \sigma_z^2 \rho_z(\|\boldsymbol{s}_{z,i} - \boldsymbol{s}_{z,j}\|_2, \|t_{z,i} - t_{z,j}\|_2), \tag{3.1}$$

where the $\boldsymbol{\mu_z}$ is the expectation, which is expanded by a set of basis functions, e.g. natural splines and $\boldsymbol{\Sigma}_z$ is the variance-covariance matrix constructed by correlation function ρ_z. In the stage, we need to generate two sets of AOD prediction at this spatiotemporal coordinates matched with $PM_{2.5}$ monitors ($\hat{\boldsymbol{z}}_y$) and at the regular spatiotemporal grid ($\hat{\boldsymbol{z}}_p$):

$$\begin{gathered}\hat{\boldsymbol{z}}_y = \hat{f}(\boldsymbol{s}_y, \boldsymbol{t}_y) + \boldsymbol{W}_{y,z}[\boldsymbol{z} - \hat{f}(\boldsymbol{s}_z, \boldsymbol{t}_z)], \\ \text{where} \quad W_{y,z} = \boldsymbol{C}_{y,z}(\widehat{\boldsymbol{\Sigma}}_z)^{-1}, \quad (\boldsymbol{C}_{y,z})_{i,j} = \widehat{\sigma_z^2}\widehat{\rho_z}(\|\boldsymbol{s}_{y,i} - \boldsymbol{s}_{z,j}\|_2, \|t_{y,i} - t_{z,j}\|_2)\end{gathered}$$

$$\begin{gathered}\hat{\boldsymbol{z}}_p = \hat{f}(\boldsymbol{s}_p, \boldsymbol{t}_p) + \boldsymbol{W}_{p,z}[\boldsymbol{z} - \hat{f}(\boldsymbol{s}_z, \boldsymbol{t}_z)], \\ \text{where} \quad W_{p,z} = \boldsymbol{C}_{p,z}(\widehat{\boldsymbol{\Sigma}}_z)^{-1}, \quad (\boldsymbol{C}_{p,z})_{i,j} = \widehat{\sigma_z^2}\widehat{\rho_z}(\|\boldsymbol{s}_{p,i} - \boldsymbol{s}_{z,j}\|_2, \|t_{p,i} - t_{z,j}\|_2).\end{gathered}$$

In stage 2, we need to associate $PM_{2.5}$ with adjusted AOD ($\hat{\boldsymbol{z}}_y$) in a varying coefficient mixed effect model as follows:

$$\begin{gathered}\boldsymbol{y} \sim N_n(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \ \boldsymbol{\mu}_y = f_1(\boldsymbol{t}_y)\hat{\boldsymbol{z}}_y + f_2(\boldsymbol{s}_y) + f_3(\boldsymbol{t}_y), \\ (\boldsymbol{\Sigma}_y)_{i,j} = \sigma_y^2 \rho_y(\|\boldsymbol{s}_{y,i} - \boldsymbol{s}_{y,j}\|_2, \|t_{y,i} - t_{y,j}\|_2).\end{gathered} \tag{3.2}$$

The final predictors of $PM_{2.5}$ mass concentration at the regular spatiotemporal grid can be constructed as

$$\begin{gathered}\hat{\boldsymbol{y}}_p = \hat{f}_1(\hat{z}_p) + \hat{f}_2(\boldsymbol{s}_p) + \hat{f}_3(\boldsymbol{t}_p) + \boldsymbol{W}_{p,y}[\boldsymbol{y} - \hat{f}_1(\hat{z}_y) - \hat{f}_2(\boldsymbol{s}_y) - \hat{f}_3(\boldsymbol{t}_y)], \\ \text{where} \quad W_{p,y} = \boldsymbol{C}_{p,y}(\widehat{\boldsymbol{\Sigma}}_y)^{-1}, \quad (\boldsymbol{C}_{p,y})_{i,j} = \widehat{\sigma_y^2}\widehat{\rho_y}(\|\boldsymbol{s}_{p,i} - \boldsymbol{s}_{y,j}\|_2, \|t_{p,i} - t_{y,j}\|_2),\end{gathered}$$

which is also known as the best linear unbiased predictor (BLUP) in mixed effects model theory. Inference of the two-step model can be performed by an iterative algorithm of estimating non-linear functions ($f(\cdot), f_1(\cdot), f_2(\cdot)$) using penalized least square estimation similar to generalized additive model [Hastie et al., 2009] and estimating covariance functions ($\sigma_z^2, \rho_z, \sigma_y^2, \rho_y$) using the variogram approach as Section 2.1.1 based on the residuals. In penalized least squares estimation we should use generalized least squares instead of simple least squares to consider spatiotemporal autocorrelation. Due to the complexity caused by

inverting a large spatiotemporal variance-covariance matrix, we apply the two-step algorithm separately for each year's data, which allows the estimated parameters to be different in different years.

To evaluate our models for the two stages, we applied cross validation methods, by leaving out a set of samples (AOD or $PM_{2.5}$) as a test dataset and using the remaining samples to train the models (space-time Kriging for AOD or varying-coefficient mixed model for $PM_{2.5}$). We separately calculated three types of cross validation:

1. *Standard 10-fold cross-validation (CV_{10})*: randomly leave out 10% of samples;
2. *Daily 10-fold cross-validation (CV_D)*: randomly select 10% of days and leave out all samples for those days;
3. *Site cross-validation ($CV_S^{(Location)}$)*: leave out all samples for one site (for $PM_{2.5}$) or within one area (for AOD).

We assessed our algorithm from different perspectives: 1) CV_{10} estimates general errors of our algorithm, 2) CV_D evaluates model performance for the case of missing measurements in one whole day (e.g., missing values of AOD caused by cloud cover) and is focused on assessing predictability in the temporal dimension, and 3) CV_S is focused on assessing spatial predictability and we construct $CV_S^{(center)}$ $CV_S^{(edge)}$ separately at the center and edge of our study domain. For each type of cross-validation, we calculated both Pearson R^2 and root mean square error (RMSE) between the predicted and measured values. Furthermore, for $PM_{2.5}$ prediction, we compared our two-step predictor with the fixed-effect-only predictor (fixed effects of varying-coefficient mixed effects model) and the random-effect-only predictor (spatiotemporal Kriging).

3.3 RESULTS

3.3.1 Descriptive Statistics: Long-term Variations of AOD and $PM_{2.5}$

We aggregated all measurements of $PM_{2.5}$ and AOD within our study domain to daily and seasonal averages and applied local regression smoothing (LOESS) [Cleveland, 1979] to ex-

tract the long-term trends. Figure 11 displays seasonal variations and slightly decreasing long-term trends. Table 5 shows the summary statistics. Both $PM_{2.5}$ and AOD tended to be highest in autumn and lowest in winter. The analysis of variance (ANOVA) shows that $PM_{2.5}$ and AOD are significantly different between different years or seasons ($P \ll 0.0001$), when ignoring temporal autocorrelations. Figure 10 also displays long-term spatial variations of $PM_{2.5}$.

3.3.2 AOD Smoothing

We estimated the product-sum variogram and the corresponding covariance function of the spatiotemporal field of AOD based on its empirical variograms (Figure 12). According to the estimated covariance function, correlation of AOD attenuated quickly in the temporal dimension and reaches its minimum by a 2-3 day lag, which confirms that the 7-day time window is wide enough to catch highly correlated values neighboring predicted coordinates. Figure 16 (a) displays some examples of spatiotemporal Kriging of AOD.

One benefit of AOD smoothing is to reduce errors caused by spatial misalignment between AOD and $PM_{2.5}$ mass concentration (Figure 13). $PM_{2.5}$ is 1.7% more highly correlated with smoothed AOD (Pearson $R^2 = 0.4466$) than raw AOD measurements nearest to monitoring stations ($R^2 = 0.4392$). Averaging AOD and $PM_{2.5}$ into monthly values can decrease random noise in their measurements; and thus can improve their correlation and highlights the benefits of spatiotemporal smoothing. After averaging, AOD smoothing increases their correlation by 21% (from $R^2 = 0.5256$ to $R^2 = 0.6317$). Ignoring spatial variations and averaging all sites' monthly means within the study domain into a single time series, correlations are further increased ($R^2 = 0.7125$ for smoothed AOD and $R^2 = 0.5794$ for nearest raw AOD), which confirmed AOD's representativeness of $PM_{2.5}$, especially in its temporal variations. More details of correlations between AOD and $PM_{2.5}$ are shown in Table 6.

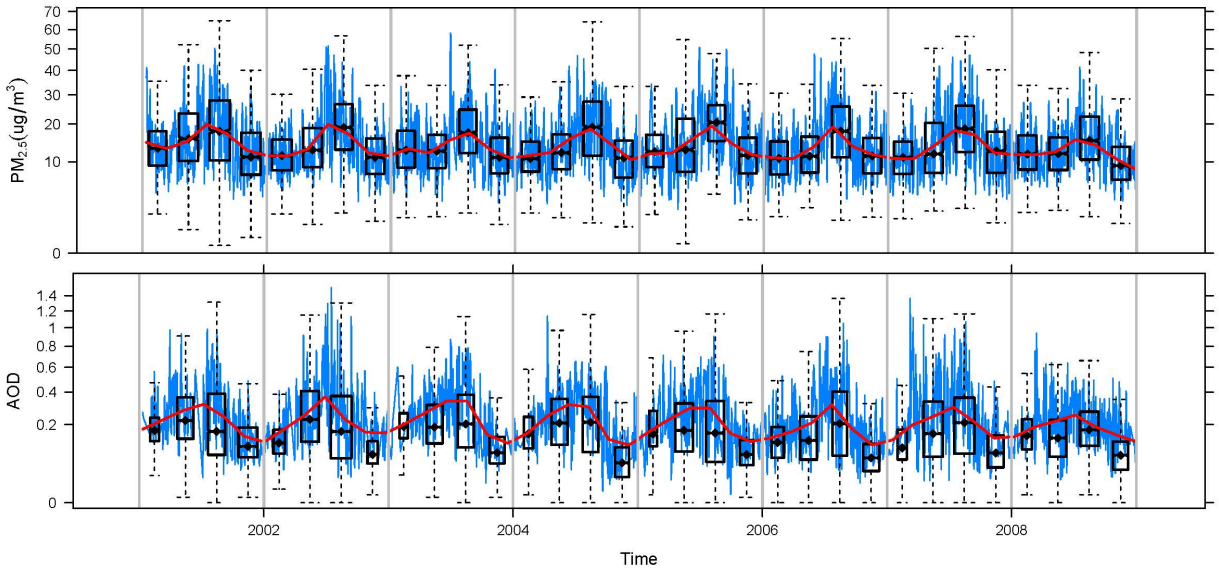
3.3.3 Correlation between AOD and $PM_{2.5}$

As some unobserved seasonally-varying factors such as relative humidity can potentially influence the association between AOD and $PM_{2.5}$ [Crumeyrolle et al., 2013], the $PM_{2.5}$ −

Table 5: Statistical summary of $PM_{2.5}$ and AOD.

	$PM_{2.5}(\mu g/m^3)$		AOD (μ)	
	Mean	SD	Mean	SD
Total	15.23	9.57	0.15	0.20
2001	16.84	10.71	0.16	0.19
2002	15.22	9.57	0.20	0.29
2003	15.27	9.41	0.16	0.18
2004	14.82	9.61	0.15	0.19
2005	16.16	10.25	0.15	0.20
2006	14.48	9.07	0.12	0.18
2007	15.22	9.38	0.16	0.21
2008	13.68	7.90	0.12	0.15
P*	$\ll 0.0001$		$\ll 0.0001$	
Spring	13.13	7.34	0.12	0.15
Summer	15.18	9.98	0.19	0.21
Autumn	19.77	10.68	0.20	0.24
Winter	12.78	8.23	0.06	0.10
P*	$\ll 0.0001$		$\ll 0.0001$	

* P-values of analysis of variance to test the null-hypotheses H_0: *$PM_{2.5}$ (or AOD) are the same level in different years (or seasons)*



Daily averaged values are denoted by blue lines with their smoothed trends (LOESS curves with smoothing parameter $\lambda = 0.05$) denoted by red lines; Seasonally summary statistics are illustrated by boxplots.

Figure 11: Time series of aggregated $PM_{2.5}$ mass concentration ($\mu g/m^3$) and AOD.

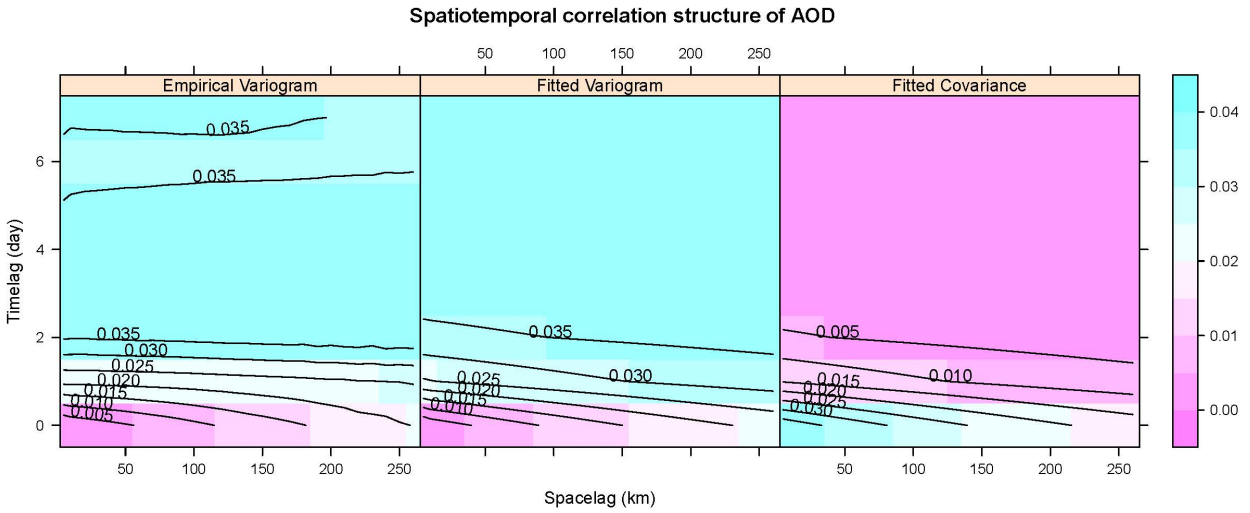
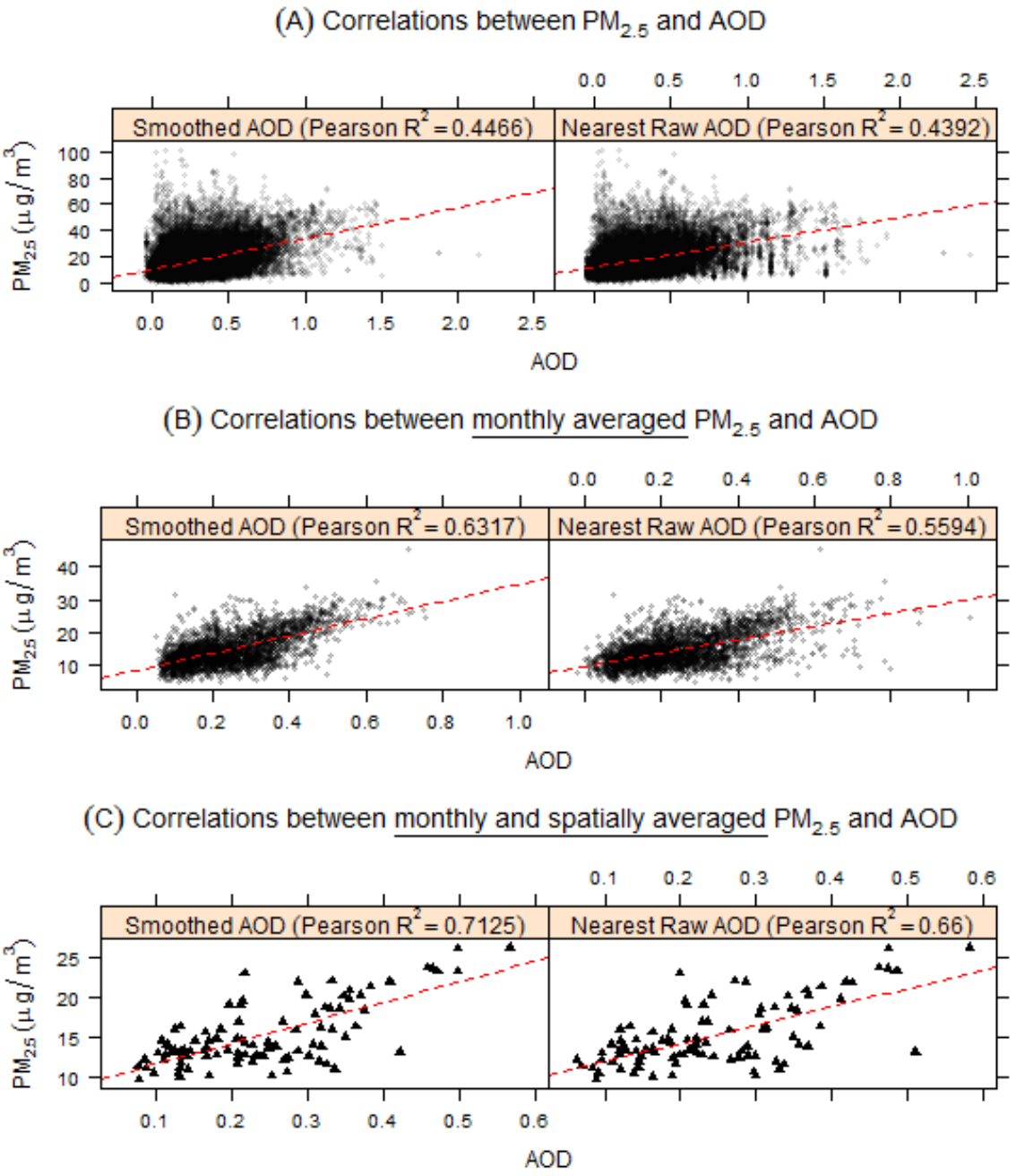


Figure 12: Empirical variograms, fitted variograms and fitted covariance functions for AOD using the product-sum structures.

AOD relationship varies periodically (Figure 14(a)). The association is stronger in the third ($R^2 = 0.5307$) and the second quarter ($R^2 = 0.4434$) but lower in fourth ($R^2 = 0.1734$) and first quarter ($R^2 = 0.1801$) as shown in Figure 14(b). The seasonally-varying correlations suggest a coefficient-varying model for predicting $PM_{2.5}$ using AOD as described in Section 3.2.

3.3.4 $PM_{2.5}$ Prediction

Figures 15 (a)-(c) display penalized least square estimates of the fixed effects for the varying-coefficient model (3.2): the temporally varying-coefficients of adjusted AOD show a regular seasonal cycle (Figure 15(a)) and suggest stronger associations between AOD and $PM_{2.5}$ in the second and third than those in first and second quarter, which is consistent with analysis of temporal variations of Pearson R^2 (Figure 14); the smoothed long-term temporal variation (Figure 15 (b)) and the smoothed static spatial variation (Figure 15 (c)) in total explain 21.3% deviance of $PM_{2.5}$, while adjusted AOD explain an additional 11.6%. The covariance functions for $PM_{2.5}$ mass concentrations were iteratively estimated using the



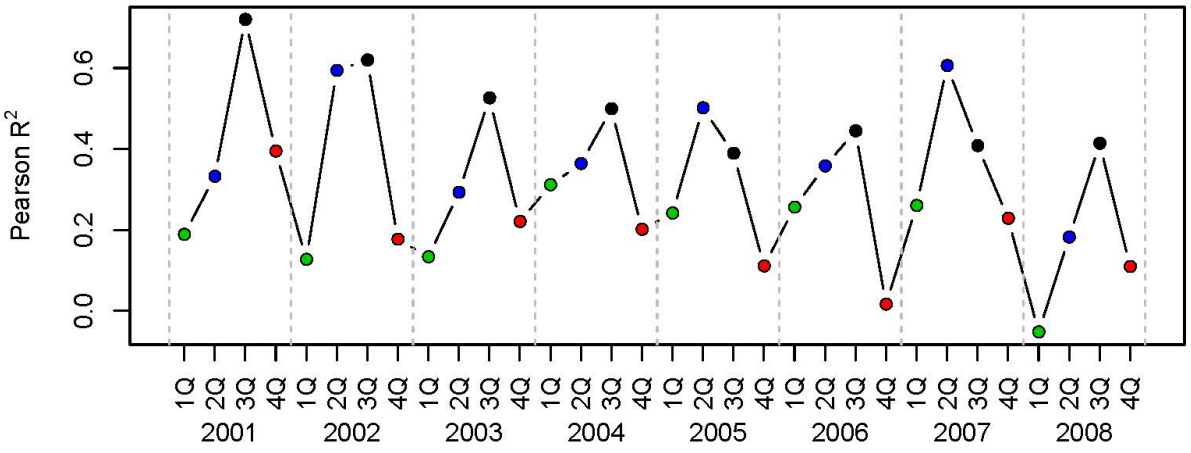
We matched each $PM_{2.5}$ mass concentration with its nearest AOD or spatiotemporal Kriging smoothed AOD at exact spatiotemporal coordinates of $PM_{2.5}$ and calculated the Pearson R^2 (A). For each monitoring site, observations are averaged monthly (B) and then the mean values of all sites are averaged (C).

Figure 13: Correlations between $PM_{2.5}(\mu g/m^3)$ mass concentration and smoothed AOD versus raw AOD in various averaged levels.

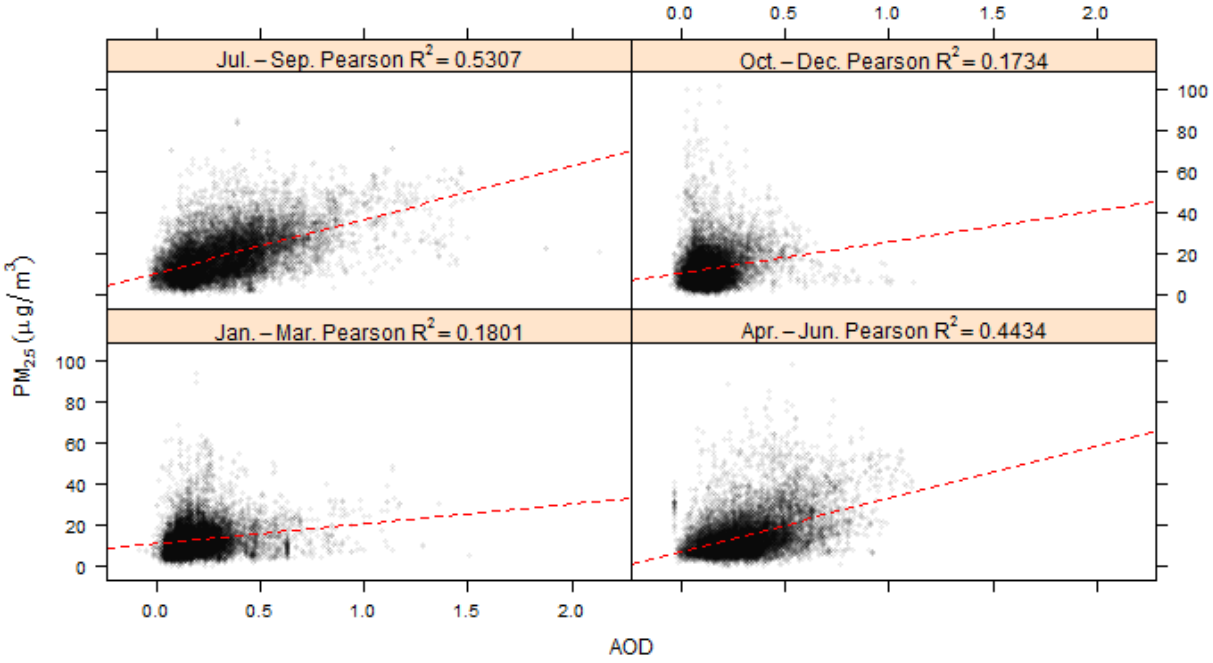
Table 6: Pearson correlation R^2s between $PM_{2.5}$ and AODs in different averaged levels.

	Pearson R^2s		
	Raw	Smoothed	
Average level		All days*	Cloudless‡
Non-aggregated	0.4392	0.4466	0.4740
Week	0.5048	0.5322	0.5456
Month	0.5256	0.6317	0.5672
Quarter	0.5912	0.6896	0.6289
Year	0.3858	0.4485	0.4116
Sites†	0.5246	0.5136	0.5500
Week+Sites†	0.5649	0.6066	0.5957
Month+Sites†	0.5794	0.7125	0.6176
Quarter+Sites†	0.7203	0.8025	0.7420
Year+Sites†	0.6831	0.8267	0.7187

† values were averaged first spatially and then temporally; * AOD values were smoothed for all days; ‡ smoothed AOD in cloudy days were excluded when correlating with $PM_{2.5}$.



(a) Pearson R^2s by seasons and years

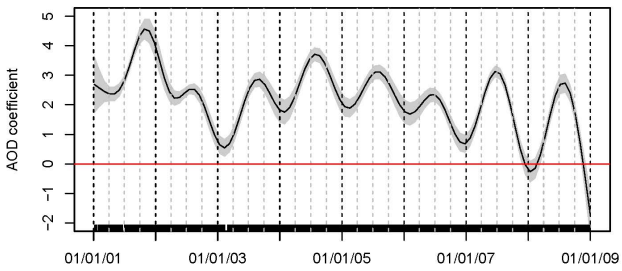


(b) Pearson R^2s by seasons

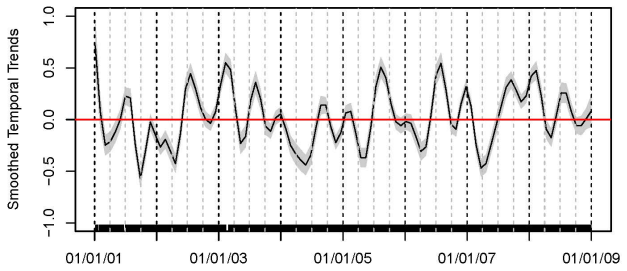
(a) shows Pearson R^2s between $PM_{2.5}$ and AOD in continuous seasons and (b) displays Pearson R^2s for four quarters, in which third quarter's R^2 (0.5307) is the highest.

Figure 14: Seasonal variations for correlations between $PM_{2.5}$ and AOD.

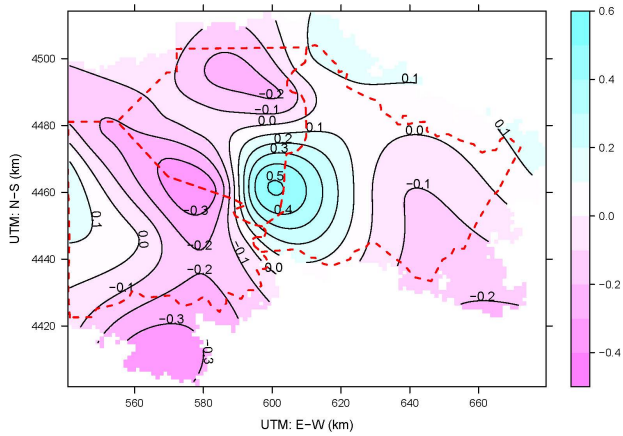
two-step algorithm mentioned in Section 3.2. Considering potentially heterogeneous spatiotemporal correlation structures, we fitted variogram and covariance functions separately for each year. Based on estimated covariance functions (Figure 15 (d)), $PM_{2.5}$ measurements are highly auto-correlated within a 4 day time lag and 50 km space lag. Examples of one week's predictions for $PM_{2.5}$ mass concentrations are shown in Figure 16 (b).



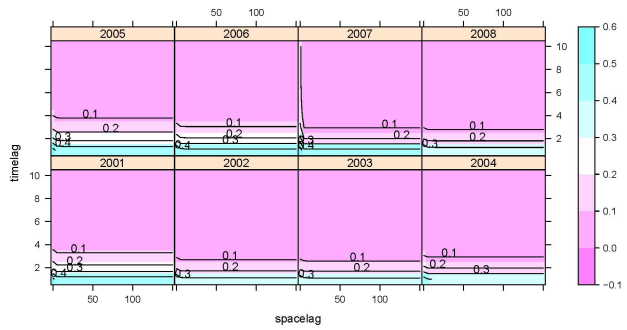
(a) Smoothed temporal varying-coefficients of AOD



(b) Smoothed long-term temporal Variation of $PM_{2.5}$



(c) Smoothed spatial variation of $PM_{2.5}$



(d) Covariance functions estimated by GLS resiudals of $PM_{2.5}$ in separate years

Figure 15: Fixed effects for varying-coefficient model associating $PM_{2.5}$ mass concentration with adjusted AOD ($f_1(\cdot), f_2(\cdot)f_3(\cdot)$) using penalized generalized least square and covariance functions estimated by variograms for GLS residuals.

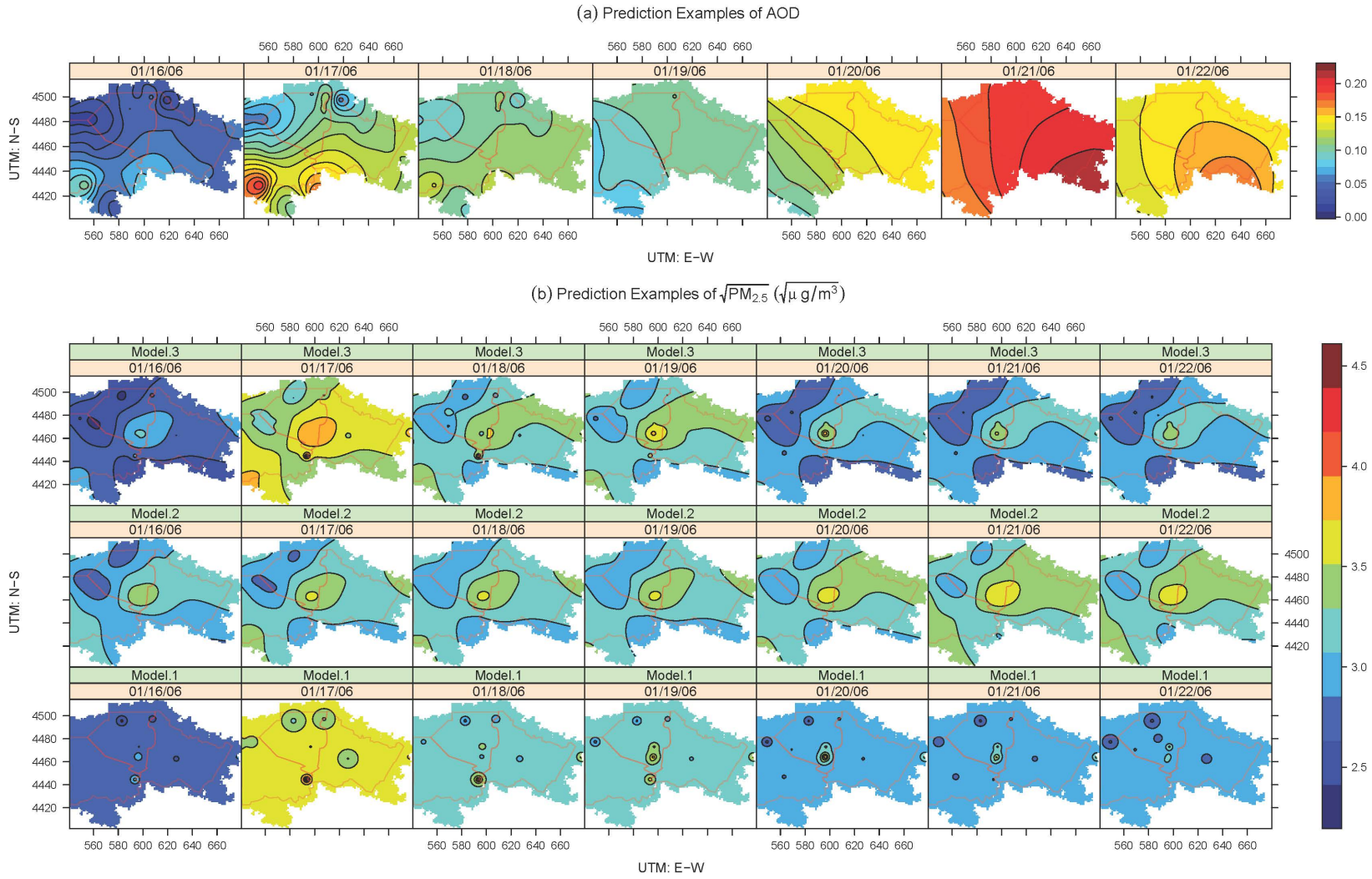


Figure 16: Examples of smoothed AOD (a) and predicted $PM_{2.5}$ mass concentrations based on three models (b).

3.3.5 Cross-validations

The cross validation results are displayed in Table 7. For each type of cross validation, the two criteria were consistent: smaller RMSE always indicates higher Pearson R^2, thus the following interpretation focuses on RMSE. For AOD, CV_S RMSEs (0.0430 and 0.0398) are close to CV_{10} RMSE (0.0313), which suggests structural missing in spatial dimension has no significant influence on AOD prediction; while CV_D RMSE (0.1878) is around 5 time higher than CV_{10} RMSE, which suggests structural missing in temporal dimension significantly increases prediction variance of AOD. In addition, due to the homogeneous spatial distribution of AOD samples, RMSEs of $CV_S^{(center)}$ and $CV_S^{(edge)}$ are close to each other. The cross validation analysis also suggests that AOD values are more predictable in spatial distribution rather than temporal variations. For $PM_{2.5}$, we construct three different predictors in cross-validation analysis:

1. **Model 1**: We only consider the random effects of spatiotemporal variations of $PM_{2.5}$ and construct the BLUP based on spatiotemporal Kriging of $PM_{2.5}$;
2. **Model 2**: We construct the varying-coefficient mixed model but only consider fixed effects in prediction;
3. **Model 3**: We construct BLUP of the varying-coefficient mixed model considering both fixed and random effects.

The performance of Model 2 is the worst (Figure 17); Comparing with Model 1, the RMSE of Model 3 is decreased by 3.0% in CV_{10}, 4.9% in CV_D and 4.2% in $CV_S^{(edge)}$ but increased 2.4% in $CV_S^{(center)}$, which suggests Model 3 in general is better than Model 1, especially in temporal prediction. In all of the three models, we find a similar pattern of CV_D and CV_S as AOD's cross validation, which indicates that temporal information loss is more harmful than spatial information loss in $PM_{2.5}$ prediction; while, unlike AOD, $CV_S^{(edge)}$ is higher than $CV_S^{(center)}$ in $PM_{2.5}$ Model 1 and Model 3 but slightly lower in Model 2, possibly because in both Model 1 and Model 3, we consider random effects in predictors, which are potentially affected by the heterogeneous spatial distribution of $PM_{2.5}$ monitoring stations.

In addition, in Section 2.1.4, the RMSE of 10-fold cross-validation for spatiotemporal Kriging of $PM_{2.5}$ using logarithm transformation was 21.20 (Table 1) which is much larger

Table 7: Summary of cross validations of $PM_{2.5}$ and AOD.

	RMSE				Pearson R^2			
	CV_{10}	CV_D	$CV_S^{(center)}$	$CV_S^{(edge)}$	CV_{10}	CV_D	$CV_S^{(center)}$	$CV_S^{(edge)}$
AOD (μ)	.0313	.1878	.0430	.0398	.9882	.5511	.9796	.9775
$\sqrt{PM_{2.5}}$ ($\mu g/m^3$): Model 1 *	.4573	.6752	.3758	.5010	.9109	.8013	.9383	.8820
$\sqrt{PM_{2.5}}$ ($\mu g/m^3$): Model 2 ‡	.9006	.9110	.8787	.8632	.5776	.5630	.5764	.5030
$\sqrt{PM_{2.5}}$ ($\mu g/m^3$): Model 3 †	.4435	.6422	.3847	.4801	.9157	.8179	.9359	.8832

For $PM_{2.5}$, the cross validation was calculated using square transformed $PM_{2.5}$ and their corresponding predictions. * Model 1 was spatiotemporal Kriging of $PM_{2.5}$; ‡ Model 2 was prediction based on only fixed effects of the varying-coefficient mixed model; † Model 3 was BLUP constructed based on both fixed and spatiotemporal random effect of the varying-coefficient mixed model.

than that for both the spatiotemporal Kriging (with squared root transformation and yearly-specific variograms) and varying-coefficient mixed effect model (Table 7) in this section.

3.4 DISCUSSION

AOD has been used as a common surrogate $PM_{2.5}$ to predict its spatiotemporal variations in previous studies [Paciorek et al., 2008, Liu et al., 2009]. Kumar et. al., (2010), however, concluded that five major components may limit the $PM_{2.5}$-AOD association: (1) aerosol types; (2) control for spatiotemporal structure in the statistical model and mismatch between AOD and $PM_{2.5}$, (3) spatiotemporal resolution; (4) collocation, and (5) integration. In our analysis, we applied smoothing of AOD to calibrate the spatial misalignment, which can reduce the mismatch between AOD and $PM_{2.5}$ in spatiotemporal resolution and integration. AOD and $PM_{2.5}$ are spatiotemporal mismatched, because $PM_{2.5}$ samples are spatial

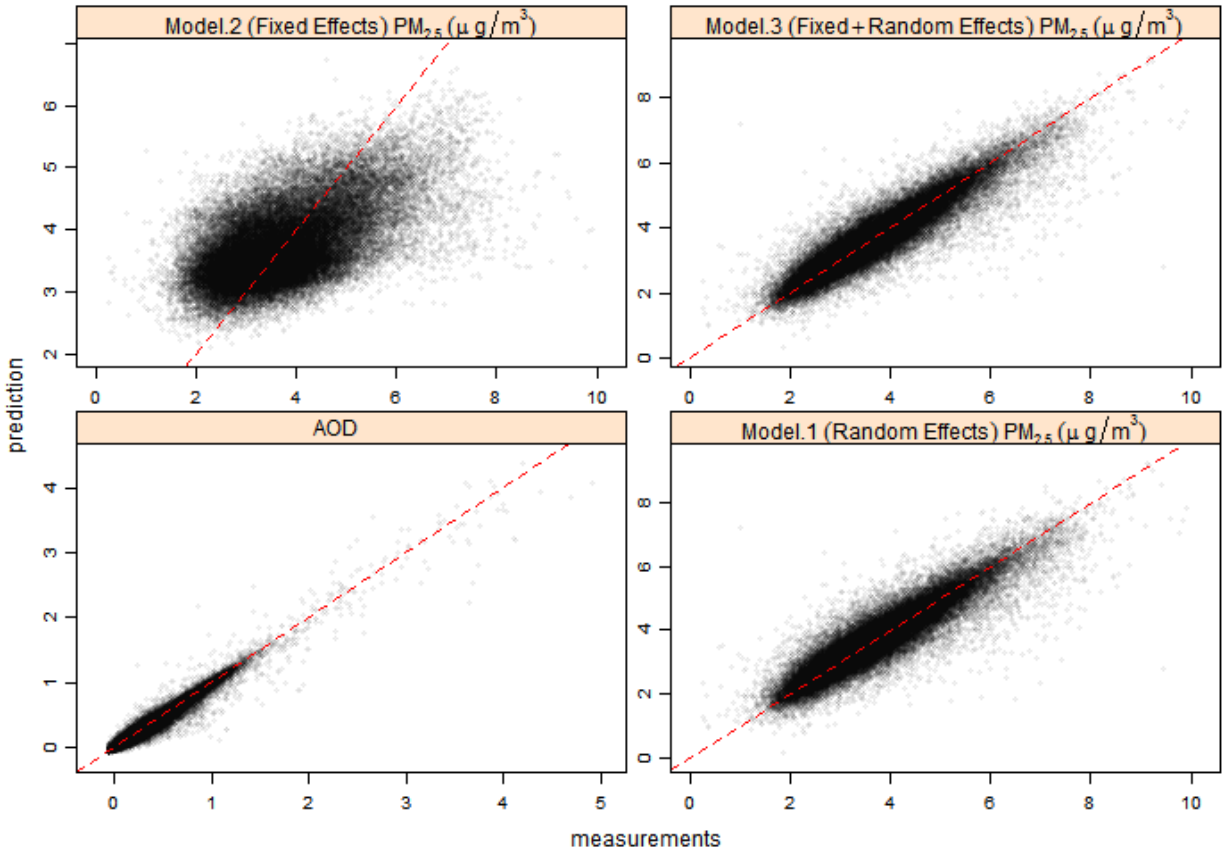


Figure 17: 10-folds cross validations for AOD and $PM_{2.5}$: Predictions vs measurements.

point measurements aggregated into daily averages, while Level 2 AOD values are temporal point measurements aggregated spatially within $10 \times 10km^2$ pixels. Smoothing AOD can exactly match centroids of satellite AOD measurements with locations of $PM_{2.5}$ stations and reconstruct the spatial trend within pixels, which are lost in AOD spatial integration.

AOD retrieval is not available on cloudy days, which also contributes to the mismatch of location. Our AOD smoothing approach may potentially fix the problem of cloud coverage but its improvements are not obvious. In order to determine the effect, we calculated Pearson correlation R^2s (Table 6) between $PM_{2.5}$ and smoothed AOD only in the cloudless days under various aggregation levels and compared them with R^2s in all days (which reflect $PM_{2.5}$-AOD association in our prediction model, as we regressed smoothed AOD with all available $PM_{2.5}$ samples). The advantages of including smoothed AOD on cloudy days into our data pool are: (1) increasing sample size, and (2) providing more information of temporal variations of AOD. The disadvantage is the increase in noise of the smoothed AOD, because predicted AOD values on cloudy days have larger errors than those on cloudless days. The results in Table 6 confirm the trade-off between the advantages and disadvantages. The R^2 for cloudless days is higher than the R^2 for all days in non-averaged data, which may reflects AOD on cloudless days has less noise; while with aggregating data in temporal dimension, the total noise in our measurements was reduced and adding smoothed AOD for cloudy days into our data pool shows more advantages than disadvantages. This is confirmed in the comparison of R^2s: in weekly data, the difference between the two R^2 was reduced and from monthly to yearly averaging, the previous trend is reversed. The comparisons of spatially averaged R^2s between data for all days and cloudless days have similar patterns.

In our statistical modeling, we assumed a flexible $PM_{2.5}$-AOD association and more generalized spatiotemporal dependence structure in case of overestimating AOD's contribution to $PM_{2.5}$ prediction. The seasonal variation of $PM_{2.5}$-AOD association has been confirmed in both our study and previous research [Paciorek et al., 2008, Zhang et al., 2009]. Paciorek and Liu applied a similar time varying coefficient modeling approach and associated AOD with $PM_{2.5}$ in the mid-Atlantic study region of the United States (covering all of our study domain) and concluded that AOD had no significant contribution to predicting $PM_{2.5}$, especially its spatial variations [Paciorek and Liu, 2008]. Paciorek and Liu predicted monthly

Table 8: Comparing our model with Paciorek and Liu's model [Paciorek and Liu, 2008].

	Our Model†		Paciorek and Liu's Model‡	
Model	R^2	RMSE	R^2	RMSE
Daily				
No AOD*	0.9102	3.8004	-	-
With AOD♯	0.9155	3.6858	-	-
Monthly				
No AOD*	0.9917	0.6336	0.794	3.22
With AOD♯	0.9743	1.0961	0.794	3.22
Yearly				
No AOD*	0.9939	0.2880	0.463	1.33
With AOD♯	0.9690	0.5348	0.467	1.32

† We transformed daily predicted $\sqrt{PM_{2.5}}$ in our 10-fold cross validation dataset back to $PM_{2.5}$, averaged in monthly and yearly levels, and calculated their correlation Pearson R^2 with measured $PM_{2.5}$ and RMSE; ‡ The results are cited from Table 2 in Paciorek and Liu's paper [Paciorek and Liu, 2008]; * No AOD model is Model 1 and ♯ AOD model is Model 3 in Section 3.3.5

and yearly $PM_{2.5}$ and evaluated their models by cross validation methods (Table 8). In order to compare our results with their findings, we averaged our CV_{10} results (Table 8). Our results are consistent with Paciorek and Liu's finding that in monthly and yearly predictions, the models including AOD do not perform better than the models without AOD. At the daily level, however, our results show that the model with AOD is better, which suggests that AOD measurements mainly predict short-term spatiotemporal variations of $PM_{2.5}$ rather than its long-term spatial trend. Even though both RMSE and R^2 shows that our model possibly is better than Paciorek and Liu's model, the RMSE may be non-comparable in different studies due to divergence in scale of air pollutants in different study domain.

We cautiously made use of AOD to predict $PM_{2.5}$ mass concentrations, but our study is still limited because we did not distinguish different types of AOD and did not control some factors that can potentially affect $PM_{2.5}$-AOD associations, such as relative humidity, planetary boundary layer height, mean sea-level pressure, and precipitation.

In statistical analysis, our approach was limited in some respects. Firstly, as the BLUPs are constructed using their neighboring measurements, the $PM_{2.5}$ predictions suffer from the sparse and non-homogeneous distribution of $PM_{2.5}$ monitoring stations, especially in Washington and Westmoreland counties. Secondly, a non-separable covariance structure that we used to capture the spatiotemporal dependence structure was flexible and reasonable compared with previous studies, but the parametric assumption is still very strict and lost some capacity to capture certain spatiotemporal characteristics, like the clustering effect of spatial dependence [Reilly and Gelman, 2007] and the periodic effect of spatiotemporal dependence [Guttorp et al., 1994]. Thirdly, when estimating Kriging weights, we ignored spatiotemporal dependence of the measurements 7 days away by manually assigning a time window of 7 days according to empirical experience in order to reduce computational loads. The subjective choice of time window introduced unknown risk into our models and likely made our predictions diverge from the optimum.

3.5 CONCLUSION

In this chapter, we explored the long-term spatiotemporal variations of both $PM_{2.5}$ and MODIS AOD, and the seasonal varying $PM_{2.5}$-AOD association in the Pittsburgh region. Using AOD, we developed a time-varying mixed effect model with a product-sum spatiotemporal covariance, which is able to predict $PM_{2.5}$ values at the un-sampled spatiotemporal coordinates in the study domain with optimal predicting errors for use in our further studies.

4.0 ESTIMATING THE SPATIAL DISTRIBUTION OF O_3 IN THE CONTINENTAL UNITED STATES FROM THE OZONE MEASUREMENT INSTRUMENT O_3 PROFILE USING A LATENT VECTOR SPATIAL MODEL

Using satellite measurements to predict spatiotemporal variations of gaseous pollutants, especially for O_3 is more challenging than using AOD to predict $PM_{2.5}$, and thus there have been few study using satellite measurements of O_3. It is easy to show that satellite measurements captures temporal variations of ground surface O_3, therefore confirming the spatial correlation between satellite and monitoring O_3 is the key problem to be solved in this chapter. In this section, we are going to explore spatial correlation between satellite and ground surface monitoring measurements of O_3 in the continental United States by constructing a spatial latent vector model.

4.1 INTRODUCTION AND DATA

4.1.1 Introduction

Ground surface ozone (O_3) is the principal component of photo-chemical air pollutants and through epidemiological studies, its inhalation exposure has been associated with adverse health outcomes including inflammatory reactions in the lung [Devlin et al., 1991], decreased functions of airways [Tager et al., 2005], asthma [McConnell et al., 2002, McDonnell et al., 1999] and chronic obstructive pulmonary disease [Anderson et al., 1997, Medina Ramón et al., 2006]. However, most of

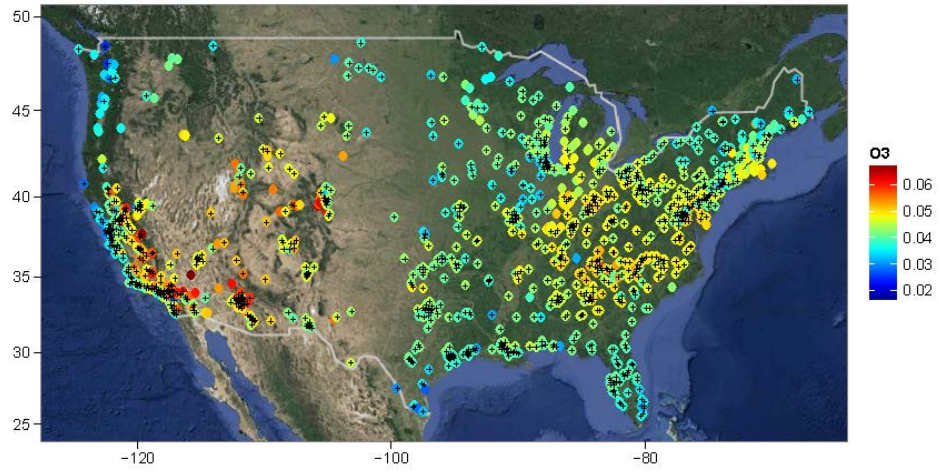
those research were designed as longitudinal studies with exposure assessment of O_3 via fixed ambient monitors and rarely took spatial variations into account [Frischer et al., 1999, Gent et al., , Loomis et al., 1996]. As O_3 is a short-lived species in the troposphere, its concentrations are determined by its mixing rate and precursors, which are highly depend on local meteorology [Dueñas et al., 2002, Pudasainee et al., 2006] and local sources (e.g., volatile organic compounds (VOCs) and nitrogen oxide (NO_x) from traffic and biomass combustion) [Zheng et al., 2009, Cohan et al., 2005]. Previous studies have reported considerable spatial variation of O_3 within urban areas [Monn, 2001], which ranged about 20% of its temporal variation [Wade et al., 2006]. Therefore estimation of the spatial variation of O_3 is critical in order to minimize exposure misclassification and to accurately evaluate the health risk of O_3, especially for cross-sectional epidemiology and large-scale cohorts [Wade et al., 2006].

Previous studies usually have estimated large-scale spatial variation of air pollutants via three methods: (1) interpolation of the monitor network using statistical models including geostatistical methods [Phillips et al., 1997], land use regression [Hoek et al., 2008] or others (e.g., Bayesian Maximum Entropy [Adam Poupart et al., 2014]); (2) atmospheric modeling (e.g., Community Multi-scale Air Quality Model (CMAQ) [Tong and Mauzerall, 2006, Liu et al.,]) and (3) satellite remote sensing [Liu et al., 2009, Van Donkelaar et al., 2006, Richter et al., 2005]). However, on the one hand, ambient monitors are often too sparsely distributed to capture detailed spatial variation of pollutants; on the other hand, atmospheric modeling or satellite remote sensing outputs may be biased [Eder and Yu, 2006, Swall and Davis, 2006] or impacted by other methodological and geophysical variables [Engel Cox et al., 2004, Martin, 2008]. In order to improve prediction, statisticians recently have designed algorithms to combine simulated outputs of atmospheric models with monitor networks through Bayesian hierarchical modeling [Fuentes and Raftery, 2005, McMillan et al., 2010]. In a similar fashion, we will make prediction using monitors fused with satellite remote sensing. In this paper, we are focusing on estimating long-term spatial variation of O_3 using a combination of the Ozone Monitor Instrument (OMI) O_3 profile and US Air Quality System (AQS) monitors.

Satellite remote sensing collects electromagnetic signals of light from the bottom to the top of the atmosphere and retrieves the vertically integrated column concentration of a trace gas from absorption of solar backscatter or emission of near-infrared light at a specific wave length corresponding to the gas [Martin, 2008]. In the stratosphere, O_3 acts as a shelter against ultraviolet radiation. In the upper troposphere O_3 acts as a greenhouse gas. O_3 only acts as air pollutants in the planetary boundary layer. Therefore unlike other trace gases (e.g. NO_2 and SO_2), which are mainly distributed in the planetary boundary layer [Martin, 2008], O_3 as an air pollutant cannot be reflected by the integrated column concentration of satellite remote sensing. However, at high atmospheric pressure, gas molecules collide with each other and the wavelength of absorption is broadened slightly (which is known as "pressure broadening") [Menzies and Chahine, 1974]. The width of the broadening is proportional to pressure and altitude; therefore, pressure broadening can be applied to retrieve the O_3 profile at a set of specific altitudes. OMI is a nadir-viewing ultraviolet-visible spectrometer launched on the Earth Observing System (EOS) Aura since July 2004 and has generated both total column O_3 and O_3 profiles [Levelt et al., 2006]. The OMI measurements at the sunlit part of its orbit are processed to generate column concentrations of O_3 for 18 layers bounded by pressure levels (surface pressure, 700, 500, 300, 200, 150, 100, 70, 50, 30, 20, 10, 7, 5, 3, 2, 1, 0.5, and 0.3 hPa) using an optimal estimation algorithm [Rodgers et al., 2000].

Previous research has shown the correlation between ground surface and free tropospheric O_3 (sea level to 3-6 km altitude) for both monthly means and maximum daily 8-h average [Jaffe, 2010], which indicates that the lowest layer (surface to ~2.5km altitude) of the OMI O_3 profile may reflect variation of ground surface O_3. In 2011, Wang et al., (2011) studied relationship between OMI O_3 profile, ozonesonde data and EPA surface monitors for August 2006 and concluded that OMI observations at the lowest layer represent the mean values of surface monitoring data and may be able to explain the larger-scale spatial variation of surface monitors. However, as satellite measurements may be influenced by a series of climate factors, including planetary boundary layer and cloud coverage, the raw measurements of O_3 of OMI are limited to represent air quality of O_3 without adjusting such factors.

In this chapter, we are going to explore the spatial correlation between yearly means of the lowest layer OMI observations and O_3 EPA AQS monitors of O_3 in continental United



The monitors included into this study are displayed with "+" symbols.

Figure 18: Locations of AQS monitors in continental United States, 2008 and their yearly averages in *ppm*.

States and construct a spatial hierarchical model to calibrate the lowest layer of the OMI O_3 profile with the AQS monitoring data and other climate variables.

4.1.2 Data Description

We first collected the maximum daily 8-h average for monitoring ozone in the the continental US taken from the EPA AQS website (http://www.epa.gov/airquality/airdata/ad_data) for the year 2008 and aggregated them into yearly averages for each monitor. We excluded the monitors with missing measurements for more than 183 days (half a year) and finally selected 1038 monitors for our study. The locations of AQS monitors and their yearly average values are displayed in Figure 18.

We collected 582,774 measurements from the Aura OMI O_3 profile [OMO3PR (V003)] collected by NASA (http://disc.sci.gsfc.nasa.gov/Aura/data-holdings/OMI) for the continental US in 2008. The spatial resolution was 13×48 km and temporal resolution is approximately

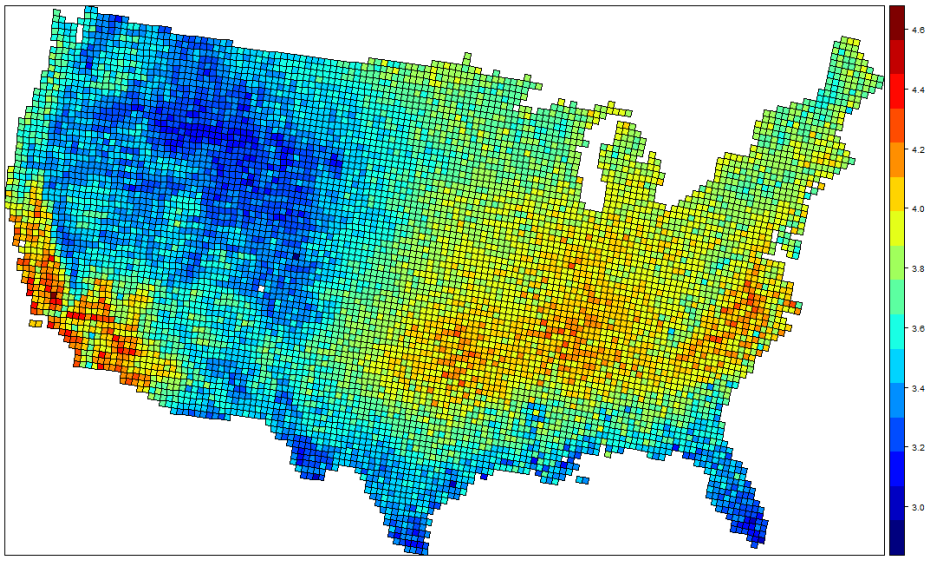


Figure 19: Yearly averages of normalized lowest layer OMI O_3 profile (DU/km) for the continental US in 2008.

once a day. We first normalized the column concentrations for the lowest layer (in Dobnson units) by the height of the column (in km) and then averaged them over a grid of approximately 32 km (projected into Lambert conformal conic system) into yearly averages ($\sim$ 75 observations for a pixel). The average height of columns for the lowest layer of the OMI O_3 profile ranged from 0.6 km to 3.1 km. We also included temperature (in K) for lowest layer from OMI profile as a covariate. The spatial variations of the normalized lowest layer OMI O_3 profile is displayed in Figure 19.

We collected a series of geographical variables including the height of planetary boundary layer, relative humidity, albedo, total cloud coverage, wind direction and wind speed as covariates to adjust the influence on satellite remote sensing from factors other than ground surface O_3 from the North American Regional Reanalysis project (NARR). NARR has generated daily measurements on a 349×277 grid of approximately 32km resolution over North American, which were also used in this study to aggregate OMI O_3 profile. Spatial variations for all the covariates are displayed in Figure 20.

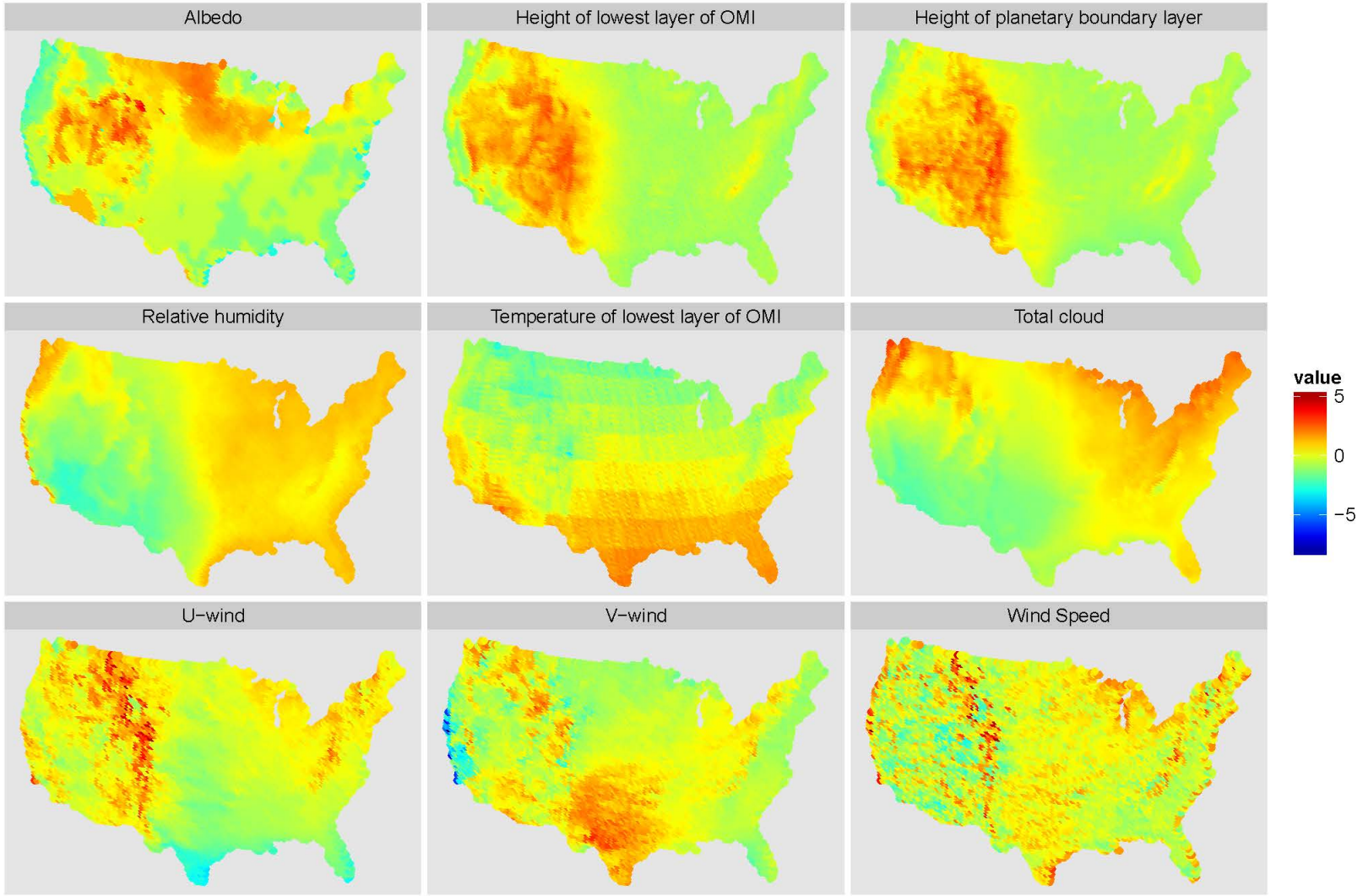


Figure 20: Yearly averages of geographical covariates for continental US in 2008. (All covariates are normalized by subtracting the mean and dividing by the standard deviation.)

4.2 STATISTICAL MODEL: LATENT VECTOR MODEL

4.2.1 Model Assumptions and Specification

In order to estimate the spatial distribution of O_3, we developed a hierarchical model with two levels to combine the satellite measurements with the monitoring data. To simplify the modeling procedure, we normalized all variables by subtracting the mean and dividing by the standard deviation. We first introduce a n-dimensional latent vector (X) with expectation of 0 to represent O_3 values that match the spatial locations of the satellite O_3 measurements (Y). In order to control the spatial autocorrelation of O_3, we add a constrain to smooth the latent vector (X). the latent vector X can be restricted using a Gaussian point process:

$$X \sim N_n(0, \mathbf{\Sigma}(d)), \text{ where } \mathbf{\Sigma}(d)_{i,j} = \sigma_x^2 exp(-\frac{d_{i,j}}{\theta})$$

or a lattice process (e.g. a spatial conditional auto-regressive (CAR)):

$$X \sim CAR(\sigma_x^2, \rho) \Rightarrow X \sim N_n(0, \sigma_x^2(D_w - \rho W)^{-1}),$$

where D_w is a $n \times n$ diagonal matrix with $(D_w)_{ii} = w_{i+}$, W is a $n \times n$ matrix with its w_{ij} elements to identify whether X_i and X_j are neighbors (if yes, $w_{ij} = 1$; else $w_{ij} = 0$) and ρ is a tuning parameter to guarantee the positive definite property of variance-covariance matrix.

In the first level of the hierarchical model, we modeled the measurement errors of m monitoring measurements $(x_1, x_2, \cdots, x_m)$. First we aligned the m measurements according to the spatial coordinates of latent vector X into a n-dimensional vector X_m, therefore, at the pixels without monitoring measurements, the elements of X_m are missing values. We modeled the measurement errors through a normal distribution:

$$X_m | X \sim N_n(X, \sigma_m^2 \mathbf{\Lambda}),$$

where $\mathbf{\Lambda}^{-1}$ is a diagonal matrix, which equals 1 at the locations with monitoring measurements and equals 0 at the location without monitors.

In the second level of the hierarchical model, we model the measurement errors of the satellite measurements Y and the influence due to other covariates Z with another linear regression model:

$$Y|X \sim N_n(X\beta + Zb, \sigma_y^2 \boldsymbol{I}).$$

As both monitoring and satellite remote sensing data are normalized to the same scale, it is reasonable to simplify the model through fixing $\beta = 1$: $Y|X \sim N_n(X + Zb, \sigma_y^2 \boldsymbol{I})$.

4.2.2 Model inference

4.2.2.1 Likelihood The inference of the hierarchical model is performed using the method of maximum likelihood. The complete likelihood can be constructed as

$$L(X, b, \sigma_x^2, \sigma_y^2, \sigma_m^2) = P(Y, X_m|X)P(X) = P(Y|X)P(X_m|X)P(X),$$

where

$$P(X) = \{(2\pi)^n \det(\sigma_x^2 \boldsymbol{\Sigma})\}^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma_x^2} X' \boldsymbol{\Sigma}^{-1} X\}$$

$$P(X_m|X) = \{(2\pi)^m \det(\sigma_m^2 \boldsymbol{\Lambda})\}^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma_m^2} (X_m - X)' \boldsymbol{\Lambda}^{-1} (X_m - X)\}$$

$$P(Y|X) = \{(2\pi)^n \det(\sigma_y^2 \boldsymbol{I})\}^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma_y^2} (Y - X - Zb)' \boldsymbol{I} (Y - X - Zb)\}.$$

We need to estimate the latent vector X and parameters $(\mu, b, \sigma_x^2, \sigma_m^2, \sigma_y^2)$ from the above likelihood. The score functions (Equation 4.1) and Fisher information matrix (Equation 4.2) for the above likelihood can be derived as follows:

$$\begin{aligned}
\partial\ell/\partial X &= 1/\sigma_y^2 \boldsymbol{I}(Y - X - Zb) + 1/\sigma_m^2 \boldsymbol{\Lambda}^{-1}(X_m - X) - 1/\sigma_x^2 \boldsymbol{\Sigma}^{-1} X \\
\partial\ell/\partial b &= Z'(Y - X - Zb)/\sigma_y^2 \\
\partial\ell/\partial \sigma_x^2 &= X' \boldsymbol{\Sigma} X/(2\sigma_x^4) - n/(2\sigma_x^2) \\
\partial\ell/\partial \sigma_m^2 &= (X_m - X)' \boldsymbol{\Lambda}^{-1}(X_m - X)/(2\sigma_m^4) - m/(2\sigma_m^2) \\
\partial\ell/\partial \sigma_y^2 &= (Y - X - Zb)' \boldsymbol{I}^{-1}(Y - X - Zb)/(2\sigma_y^4) - m/(2\sigma_y^2).
\end{aligned} \tag{4.1}$$

Setting $\partial\ell/\partial\theta = 0$:

$$\begin{aligned}
\hat{X} &= (1/\sigma_x^2\boldsymbol{\Sigma}^{-1} + 1/\sigma_m^2\boldsymbol{\Lambda}^{-1} + 1/\sigma_y^2\boldsymbol{I})^{-1}[1/\sigma_m^2\boldsymbol{\Lambda}^{-1}X_m + 1/\sigma_y^2(Y - Zb)] \\
\hat{b} &= (Z'Z)^{-1}Z'(Y - X) \\
\hat{\sigma_x^2} &= X'\boldsymbol{\Sigma}X/n \\
\hat{\sigma_m^2} &= (X_m - X)'\boldsymbol{\Lambda}^{-1}(X_m - X)/m \\
\hat{\sigma_y^2} &= (Y - X - Zb)'\boldsymbol{I}(Y - X - Zb)/n.
\end{aligned}$$

$$\begin{aligned}
I(\theta) &= \frac{\partial^2(-\ell)}{\partial\theta\partial\theta} \\
&= \begin{bmatrix}
\frac{\partial^2(-\ell)}{\partial X\partial X} & \frac{\partial^2(-\ell)}{\partial X\partial b} & \frac{\partial^2(-\ell)}{\partial X\partial\sigma_x^2} & \frac{\partial^2(-\ell)}{\partial X\partial\sigma_m^2} & \frac{\partial^2(-\ell)}{\partial X\partial\sigma_y^2} \\
\frac{\partial^2(-\ell)}{\partial b\partial X} & \frac{\partial^2(-\ell)}{\partial b\partial b} & \frac{\partial^2(-\ell)}{\partial b\partial\sigma_x^2} & \frac{\partial^2(-\ell)}{\partial b\partial\sigma_m^2} & \frac{\partial^2(-\ell)}{\partial b\partial\sigma_y^2} \\
\frac{\partial^2(-\ell)}{\partial\sigma_x^2\partial X} & \frac{\partial^2(-\ell)}{\partial\sigma_x^2\partial b} & \frac{\partial^2(-\ell)}{\partial\sigma_x^2\partial\sigma_x^2} & \frac{\partial^2(-\ell)}{\partial\sigma_x^2\partial\sigma_m^2} & \frac{\partial^2(-\ell)}{\partial\sigma_x^2\partial\sigma_y^2} \\
\frac{\partial^2(-\ell)}{\partial\sigma_m^2\partial X} & \frac{\partial^2(-\ell)}{\partial\sigma_m^2\partial b} & \frac{\partial^2(-\ell)}{\partial\sigma_m^2\partial\sigma_x^2} & \frac{\partial^2(-\ell)}{\partial\sigma_m^2\partial\sigma_m^2} & \frac{\partial^2(-\ell)}{\partial\sigma_m^2\partial\sigma_y^2} \\
\frac{\partial^2(-\ell)}{\partial\sigma_y^2\partial X} & \frac{\partial^2(-\ell)}{\partial\sigma_y^2\partial b} & \frac{\partial^2(-\ell)}{\partial\sigma_y^2\partial\sigma_x^2} & \frac{\partial^2(-\ell)}{\partial\sigma_y^2\partial\sigma_m^2} & \frac{\partial^2(-\ell)}{\partial\sigma_y^2\partial\sigma_y^2}
\end{bmatrix}
\end{aligned} \tag{4.2}$$

Applying the Bayesian rule:

$$P(X|X_m, Y) \propto P(Y, X_m|X)P(X)$$

$$\propto \exp\{-\frac{1}{2\sigma_x^2}X'\boldsymbol{\Sigma}^{-1}X - \frac{1}{2\sigma_m^2}(X_m - X)'\boldsymbol{\Lambda}^{-1}(X_m - X) - \frac{1}{2\sigma_y^2}(Y - X - Zb)'\boldsymbol{I}(Y - X - Zb)\}$$

As the above equation is a quartic form of X, then we can then conclude that posterior distribution of X is a multivariate normal distribution:

$$X|X_m, Y \sim N_n(\hat{X}, \{\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I}\}^{-1}) \tag{4.3}$$

$$\text{where } \hat{X} = (\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})^{-1}[\frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1}X_m + \frac{1}{\sigma_y^2}\boldsymbol{I}(Y - Zb)].$$

Thus we know the estimation of the latent vector X is constructed from two parts: (1) the monitoring data X_m and (2) the calibrated satellite data $(Y - Zb)$, and their weights are determined by the variances of three components: (1) the monitoring measurement error σ_m^2, (2) the satellite measurement error σ_y^2, and (3) the smoothness σ_x^2. However, estimating the

n-dimensional latent vector X may highly bias the estimation of the variance components [Pawitan, 2001], so we need to modify the log-likelihood:

$$Q = \log L(\hat{X}, \hat{b}) - \frac{1}{2}\text{logdet}[I(\hat{X})],$$

where $I(\hat{X}) = \{\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I}\}^{-1}$ is Fisher information matrix of the latent vector $\hat{X}$. We design an EM-algorithm to optimize the objective function Q as described in next section.

4.2.2.2 EM-algorithm

In the algorithm, we are going to update the latent vector X in the E-step 4.4 and update the other parameters $(b, \sigma_x^2, \sigma_m^2, \sigma_y^2)$ in the M-step.

E-step:

$$Q(\theta|\theta^{(t)}) = E_{X|Y,X_m,Z,b,\sigma_x^2,\sigma_m^2,\sigma_y^2}\big\{\ell(X, b, \sigma_x^2, \sigma_m^2, \sigma_y^2; Y, X_m, Z) - \frac{1}{2}\text{logdet}[I(X)]\big\}$$

The complete log-likelihood $(\ell(X, b, \sigma_x^2, \sigma_m^2, \sigma_y^2; Y, X_m, Z))$ consists of both linear and quadratic form of X. In order to calculate $E_{X|Y,X_m,Z,b,\sigma_x^2,\sigma_m^2,\sigma_y^2}\big[\ell(X, b, \sigma_x^2, \sigma_m^2, \sigma_y^2; Y, X_m, Z)\big]$, we can replace X in its linear forms with $\hat{X}$ and replace the quadratic form $X'(\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})X$ with $E_{X|Y,X_m,Z,b,\sigma_x^2,\sigma_m^2,\sigma_y^2}\big[X'(\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})X\big]$, and applying Equation 4.3, we can conclude that

$$X'(\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})X \sim \chi_n^2\Big(\lambda = \hat{X}'(\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})\hat{X}/2\Big)$$

$$E_{X|Y,X_m,Z,b,\sigma_x^2,\sigma_m^2,\sigma_y^2}\big[X'(\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})X\big] = 2\lambda + n \propto \hat{X}'(\frac{1}{\sigma_x^2}\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma_y^2}\boldsymbol{I})\hat{X}.$$

Accordingly, to calculate $Q(\theta|\theta^{(t)})$, we just need to replace all X terms in ℓ with $\hat{X}$, thus

$$Q(\theta|\theta^{(t)}) = -1/2\text{logdet}[I(\hat{X})] + \ell\left(\hat{X}, b^{(t)}, (\sigma_x^2)^{(t)}, (\sigma_x^2)^{(t)}, (\sigma_x^2)^{(t)}; Y, X_m, Z\right) \tag{4.4}$$

$$\hat{X} = \left\{(\sigma_x^2)^{(t)}\boldsymbol{\Sigma}^{-1} + 1/(\sigma_m^2)^{(t)}\boldsymbol{\Lambda}^{-1} + 1/(\sigma_y^2)^{(t)}\boldsymbol{I}\right\}^{-1}\left\{1/(\sigma_m^2)^{(t)}\boldsymbol{\Lambda}^{-1}X_m + 1/(\sigma_y^2)^{(t)}(Y - Zb^{(t)})\right\}$$

M-step:

$$\theta^{(t+1)} = \left[b^{(t+1)}, (\sigma_x^2)^{(t+1)}, (\sigma_m^2)^{(t+1)}, (\sigma_y^2)^{(t+1)}\right]' = \underset{\theta^{(t+1)}}{\text{argmax}} \quad Q(\theta|\theta^{(t)}).$$

In the M-step, maximizing $Q(\theta|\theta^{(t)})$ usually requires complicated computing efforts to invert large matrices, therefore the EM-algorithm is not appropriate for high-dimensional data.

4.2.3 Non-parametric Model and Convex Optimization

In order to avoid the computing burden in the EM-algorithm, we develop a non-parametric model instead of the hierarchical model and perform model inference using convex optimization. Similarly to the hierarchical model, we assume a latent vector X to represent the true O_3 in the regular grid. To simultaneously minimize measurement errors in the monitoring O_3 and satellite OMI observations and smoothness of the latent vector X, we can derive an optimization problem as follows:

$$\begin{aligned} \underset{X}{\text{argmin}} \quad & \|Y - Zb - X\|_2 \\ \text{subject to} \quad & \|X - X_m\|_2 \leq s_1 \\ & X'\boldsymbol{\Sigma}^{-1}X \leq s_2, \end{aligned} \tag{4.5}$$

where the three quartic terms correspond to the three multivariate normal distributions in the likelihood of the hierarchical model. To save computing effort in a further step, we can use a SAR model instead of a CAR model to control the smoothness in latent vector X, so that the optimization can be transformed as

$$\begin{aligned} \underset{X}{\text{argmin}} \quad & \|Y - Zb - X\|_2 \\ & +\lambda_1\|X - X_m\|_2 \\ & +\lambda_2\|(\boldsymbol{I} - \boldsymbol{C})X\|_2, \end{aligned} \tag{4.6}$$

where $\boldsymbol{C}$ is a sparse matrix of neighboring weights as illustrated in Section 2.2.1. Therefore the model inference is a quadratically constrained quadratic program (QCQP) problem and can be performed using the `cvx` [CVX Research, 2012] package in MATLAB. The tuning parameters λ_1 and λ_2 can be selected using cross-validation. The non-parametric model provides a flexible strategy to combine monitoring and satellite O_3 and we can replace the L-2 norms in Equation 4.6 with other types of norm for different potential loss functions. For example, we can use L-1 norm instead of L-2 norm to smooth the latent vector X to guarantee neighboring pixels share exactly the same value:

$$\begin{aligned} \underset{X}{\text{argmin}} \quad & \|Y - Zb - X\|_2 \\ & +\lambda_1\|X - X_m\|_2 \\ & +\lambda_2\|(\boldsymbol{I} - \boldsymbol{C})X\|_1. \end{aligned} \tag{4.7}$$

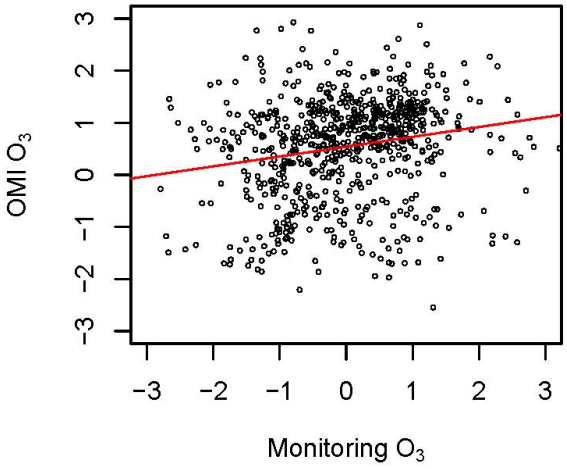
4.3 RESULTS

4.3.1 Correlation between Satellite OMI O_3 and Monitoring O_3

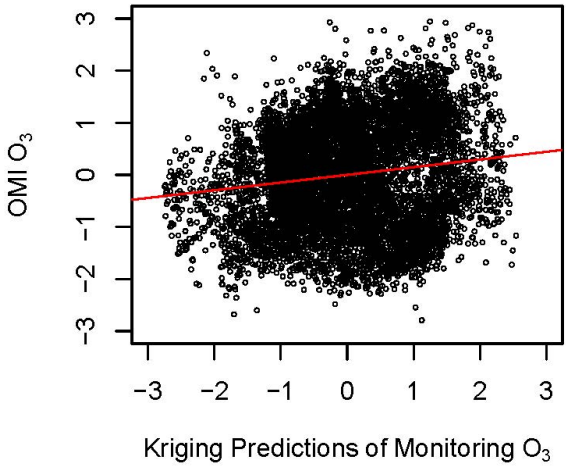
The spatial pattern map of OMI O_3 (Figure 19) captures the hot-spots of ground surface O_3 in the state of California and the mid-eastern US, but underestimates O_3 in the Rocky mountain areas. The scatterplot between unadjusted satellite O_3 of the lowest layer of OMI and ground surface monitoring O_3 is displayed in Figure 21(a) and their correlation is as low as 0.1946. In order to involve the un-monitored points into correlation analysis, we first applied the Kriging method to interpolate monitoring O_3 by the regular grid (Figure 22(b)) and then correlated the smoothed values of monitoring O_3 with OMI measurements as shown in Figure 21(b). To calibrate OMI measurements, we applied a simple regression model to associate OMI data with the height of the lowest layer of the OMI measurements, the height of the planetary boundary layer, the relative humidity, the albedo, the total cloud cover, wind direction and wind speed and correlated its residuals (Figure 22(a)) with monitoring or smoothed monitoring O_3 as displayed in Figures 21(c) and (d). Thesis figures show the increase of OMI's correlations after calibration. According to the correlation analysis, we can conclude that OMI is only representative of the spatial pattern of ground surface O_3 after calibration of other influential atmospheric factors.

4.3.2 Tunning Parameters Selection and Non-parametric Modeling Results

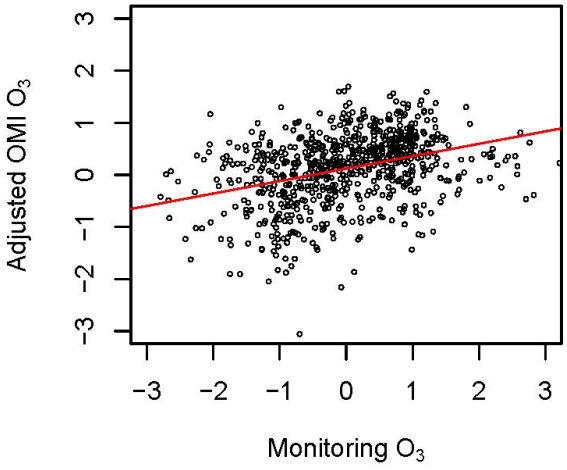
We assigned a set of fixed values (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) to the tuning parameters λ_1 and λ_2 and selected the combination that minimized the root-mean-square error (RMSE) for 10-fold cross-validation. For models 4.6 and 4.7, the curves of RMSE by tuning parameters (λ_1,λ_2) are shown in Figure 23. The optimal RMSEs for L2 and L1 model (Equations 4.6 and 4.7) 0.6552 and 0.6503 compared to 0.6557 for that of Ordinary Kriging, which reflects that both of our hierarchical latent vector models outperform the Kriging method, if only slightly.



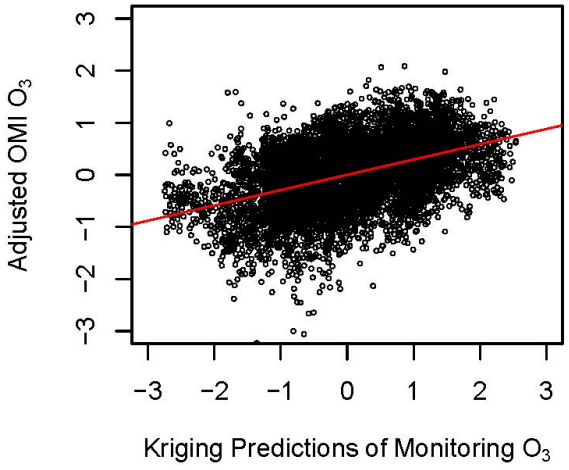
(a) OMI vs monitoring O_3 (regression line: $Y = 0.1888X + 0.5413$; $R^2 = 0.1946$)



(b) OMI vs Kringing monitoring O_3 (regression line: $Y = 0.1478X - 0.000$; $R^2 = 0.1477$)



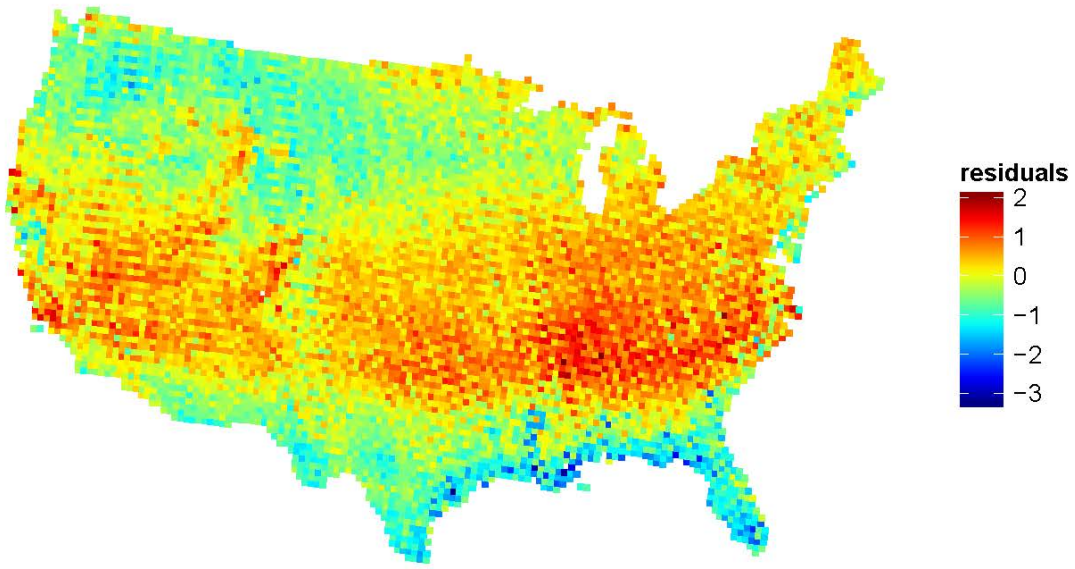
(c) Adjusted OMI vs monitoring O_3 (regression line: $Y = 0.2387X + 0.1191$; $R^2 = 0.3494$)



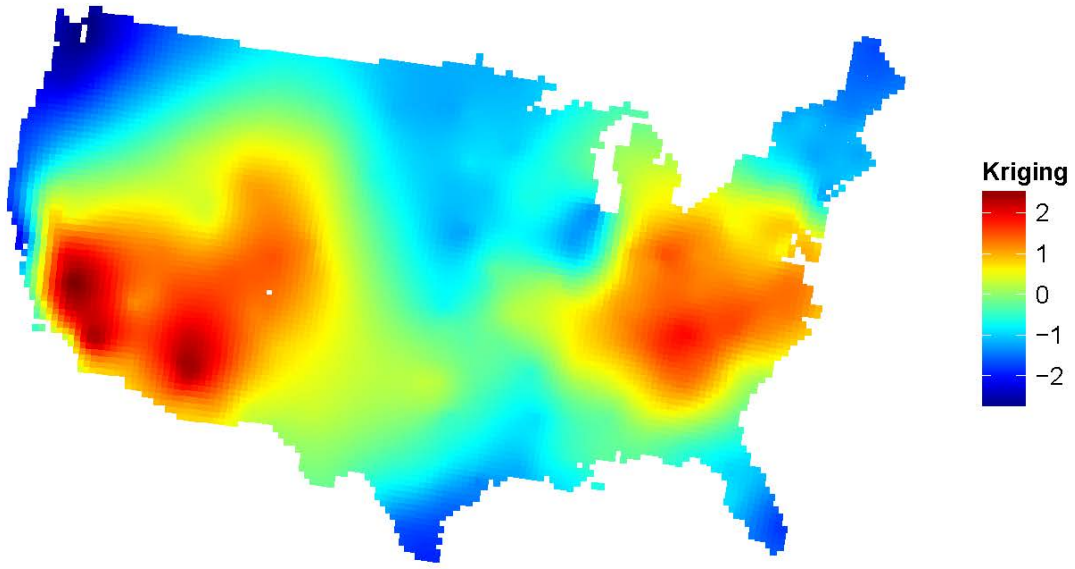
(d) Adjusted OMI vs Kriging monitoring O_3 (regression line: $Y = 0.2937X - 0.000$; $R^2 = 0.4613$)

In each figure, the red lines show simple regression lines.

Figure 21: Scatterplots between adjusted or unadjusted satellite O_3 and monitoring O_3 or Kriging smoothing of monitoring O_3 and their Pearson correlations.

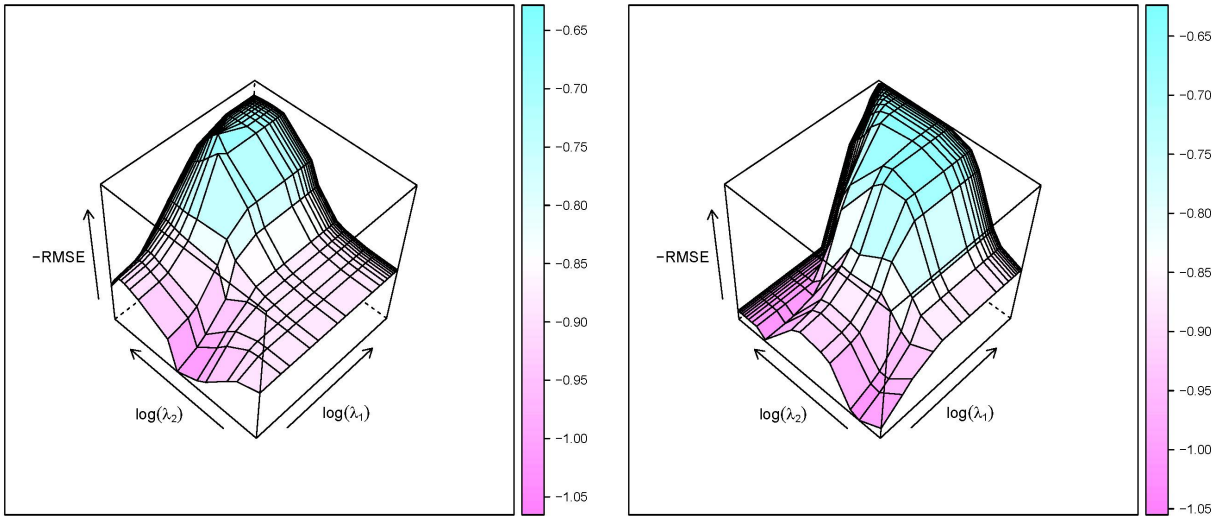


(a) Residuals of calibration model of OMI O_3



(b) Kriging of monitoring O_3

Figure 22: Mapping residuals of the calibration regression model of OMI O_3 (upper) and Kriging interpolated monitoring O_3 (lower) in the continental US.



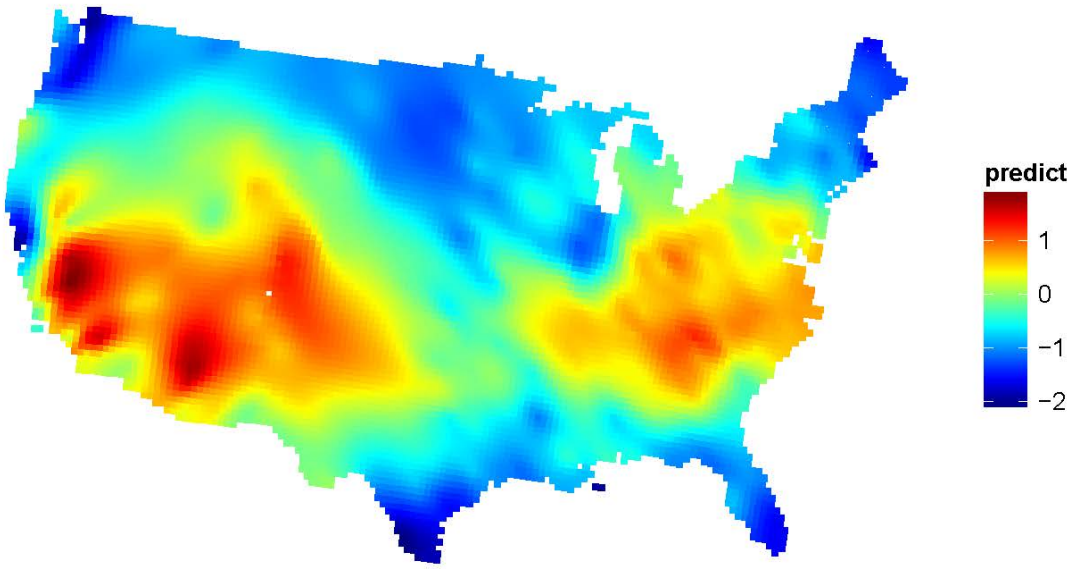
(a) L-2 spatially smoothing (model 4.6)

(b) L-1 spatially smoothing (model 4.7)

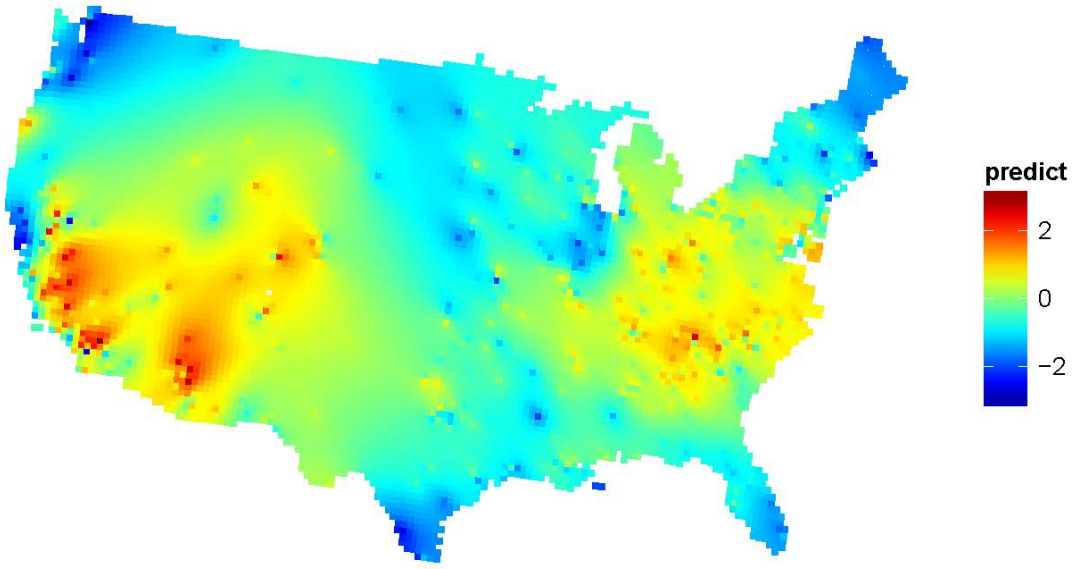
Figure 23: Tuning parameters selection: RMSE surfaces by tuning parameters, λ_1 and λ_2 for L-2 and L-1 spatially smoothing model.

4.3.3 Comparing Interpolation of Non-parametric Models with Kriging

The optimally interpolated ground surface O_3 $(\hat{X})$ by models 4.6 and 4.7 are shown in Figure 24. Comparing result of L-2 smoothing model with that of Kriging, the former map captures more spatial variation, especially in the state of California and the mid-eastern US but over-smoothed some of the extreme values, which may be because of over-smoothness in OMI measurements (as shown in Figure 21), while the L-1 smoothing model avoided this weakness and therefore is the best among the three methods according to the 10-fold cross-validation.



(a) Interpolated O_3 of L-2 spatially smoothing (model 4.6)



(b) Interpolated O_3 of L-1 spatially smoothing (model 4.7)

Figure 24: Interpolated ground surface O_3 using combination of satellite OMI and AQS monitoring O_3 by L-2 or L-1 spatially smoothing model.

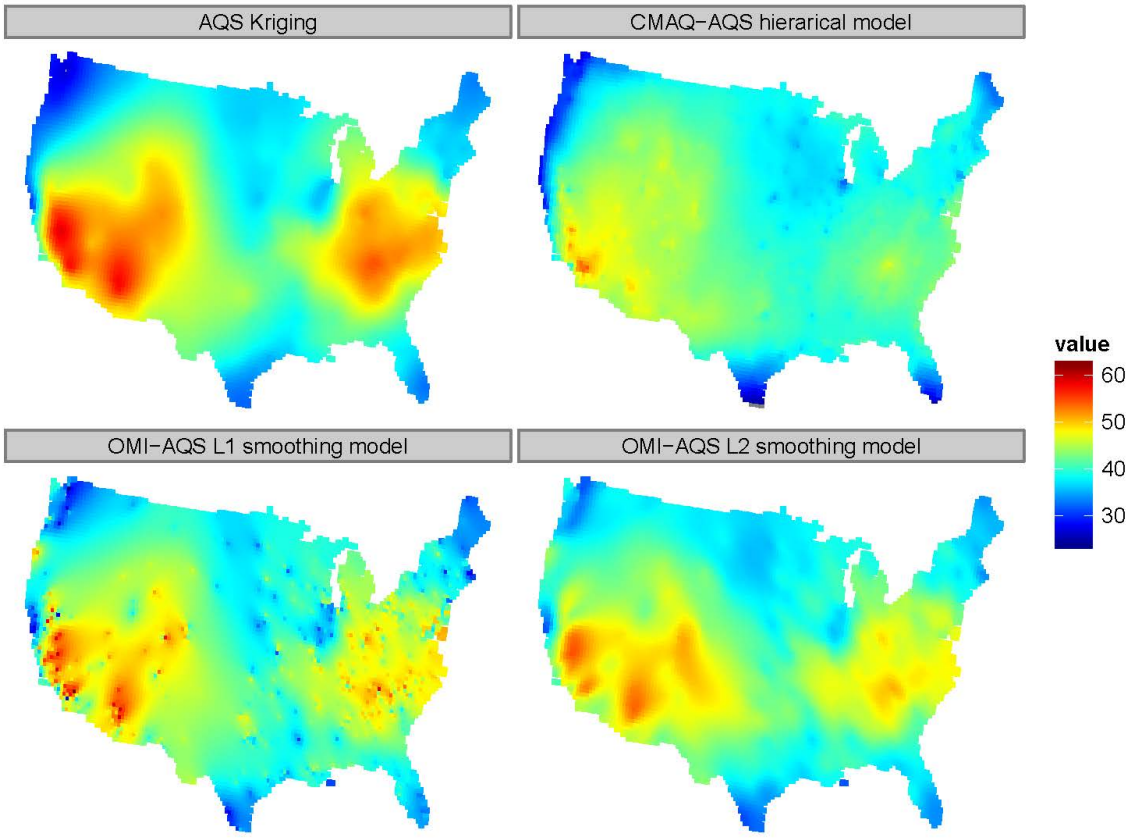


Figure 25: Estimated spatial patterns of ground surface O_3 using combinations of CMAQ, OMI and AQS data in 2008.

Table 9: Predicting accuracy of ground surface O_3 using four methods: comparing four methods' prediction with yearly averages of CASTNET monitors in 2008.

Model	AQS Kriging	OMI−AQS L1 smoothing	OMI−AQS L2 smoothing	CMAQ−AQS hierarchical model
Pearson R^2	0.7013	0.8098	0.7416	0.8902
RMSE	6.1166	4.4846	4.7505	3.7241
Biasness	4.7295	3.3043	3.7630	2.8206

4.4 DISCUSSION

In this paper, we explored the spatial relationship between the lowest layer of the OMI O_3 profile and AQS monitoring O_3. Our results reflect that satellite OMI is less predictive for the spatial pattern of the ground surface O_3 without calibration using other atmospheric variables. Even though the calibrated OMI data is highly correlated with monitoring O_3, we should not over-interpret their improvements in predicting the spatial trend of ground surface O_3 as the combined estimators from our hierarchical models only sightly decrease cross-validation errors compared with the Kriging method. Our results suggest that including satellite O_3 cannot increase predicting accuracy of ground surface O_3 significantly, but captures more locally spatial variations particularly for the L-1 smoothing model (4.7).

Another widely used ground surface O_3 estimator was calculated by combining CMAQ modeling values and AQS monitors using a spatiotemporal hierarchical model [McMillan et al., 2010]. We first averaged the daily spatiotemporal estimators of CMAQ-AQS hierarchical model into yearly values by the fixed grid and compared the result with those of our OMI-AQS hierarchical models and Kriging of AQS data as displayed in Figure 25. In order to compare the performance of the four methods, we introduce a set of external monitoring data of yearly averages of daily 8-h max O_3 from 77 sites

of CASTNET (http://epa.gov/castnet/javaweb/index.html). Through comparing four method's estimation with annual averages of observations from CASTNET's monitors, we can conclude that our OMI-AQS estimators outperform Kriging of AQS monitors but are not as accurate as CMAQ-AQS hierarchical spatiotemporal model. However, the comparisons are not able to illustrate the performance between our non-parametric hierarchical latent vector model and the Bayesian hierarchical spatiotemporal model, as the two models are applied on different scales of O_3 data. The latter model was applied to the daily measurements of AQS and CAMQ O_3, so that the study of McMillan et al., (2010) had a much larger sample size than our spatial study and captured more detailed variations of ground surface O_3.

Even though our non-parametric method avoided inverting large variance-covariance matrix in model inference compared with the EM-algorithm for the parametric model, it is limited in assessing the uncertainty in prediction. However, our non-parametric method is not restricted in our statistical analysis but provides a general framework to combine air pollutants measurements from different sources. For example, in order to further improve predicting accuracy of ground surface O_3, we could include CMAQ modeled values combined with satellite measurements and routine monitors by adding another constraint to minimize the norms between the latent vector of true O_3 and CMAQ values in our future study. In addition, we could extend the non-parametric model from space to space-time by adding another smoothing constraint of the L-1 or L-2 norm in temporal dimension.

4.5 CONCLUSION

In this section we explored the relationship between ground surface O_3 and the lowest layer of the satellite OMI O_3 profile and concluded that satellite O_3 is only predictive for the spatial pattern of ground surface O_3 after calibration of other potentially influential atmospheric factors. We also developed a latent vector hierarchical model and a non-parametric optimization for the model to combine OMI and AQS O_3. Even though our OMI-AQS estimators are not as accurate as CAMQ-AQS estimators [McMillan et al., 2010], they are

have been shown better than Kriging of AQS O_3 and capture more local spatial variation of ground surface O_3.

5.0 ASSOCIATING MORTALITY WITH AIR POLLUTANTS FROM 1999 TO 2009 IN THE PITTSBURGH REGION USING SPATIOTEMPORAL GENERALIZED ESTIMATING EQUATIONS

Air pollutants have been associated with mortality risks of cardiovascular and respiratory diseases by many epidemiological studies. However, most of these studies are longitudinally or spatially (or ecologically) designed but do not take the complex autocorrelation of health outcomes into account. In this section, we are going to develop a parameter driven spatiotemporal regression model and apply generalized estimating equations (GEEs) and a vector autoregressive (VAR) process to estimating the coefficients in the model. Compared with existing spatiotemporal methods, the algorithms described in this section avoid inverting large covariance matrices and MCMC simulation for model inference.

5.1 INTRODUCTION AND DATA

5.1.1 Introduction

5.1.1.1 Review of Epidemiology of Air Pollutants and Their Study Designs

Air pollutants have been known to be associated with adverse health outcomes since 1980s [Dockery et al., 1982, Dockery et al., 1989, Pope 3rd, 1989]. More and more epidemiological studies have revealed detailed relationships between various air pollutants including particulate matter (PM_{10} and $PM_{2.5}$), nitrogen oxides (NO, NO_2, NO_x), sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO) and so forth with both chronic and acute health effects including mortality [Pope III et al., 1992, Pope et al., 1996, Pope 3rd et al., 1999],

hospital admissions [Schwartz and Morris, 1995, Schwartz, 1996, Schwartz, 1999], specific symptoms of circulatory and respiratory diseases [Brook et al., 2004, Peters et al., 2001, Pope et al., 2006, Roemer et al., 1993, Brauer et al., 2002], and other effects such as birth defects [Ritz et al., 2000, Ritz et al., 2002].

Although many epidemiological studies have focused on air pollutants, most had designs that were either purely temporal (e.g. population-based time-series [Pope III et al., 1991], case-crossover [Levy et al., 2001], cohort [Hoek et al., 2002] and panel-based [Pope 3rd et al., 2004] studies) or purely spatial (e.g. ecological [Coyle et al., 2006], cross-sectional [Lindgren et al., 2009] and spatial case-control [Tonne et al., 2007] studies) designed studies. Previously, temporal studies have been widely used but might ignore spatial variance of air pollutants, especially in urban area, which has been shown to be comparable with temporal variance [Jerrett et al., 2005, Monn, 2001] and may bias epidemiological results due to potential exposure misclassification [Kim et al., 2005]. In recent studies, researchers have introduced spatial information into temporal designs, for example, the geographic area matched case-crossover study [Zanobetti and Schwartz, 2005].

However, either purely temporal or purely spatial designs will not account for complicated autocorrelation and may bias model inference of epidemiological studies. Statisticians have developed sophisticated models to deal with temporally correlated data (which is also known as longitudinal data) using generalized estimating equations [Zeger et al., 1988, Diggle et al., 2002] or mixed effects models [Jørgensen et al., 1996, Verbeke and Molenberghs, 2009]. While for spatially correlated data, the covariance structure of the latent stochastic process is much more complex than that of purely longitudinal data. Recently, statisticians have mainly applied hierarchical models [Zhu et al., 2003] or mixed model with complex assumptions for the random effects [Pope III et al.,] (which are similar to Bayesian hierarchical models) to deal with spatially correlated data. Thus, spatial and temporal autocorrelations have rarely been considered simultaneously in epidemiological studies of air pollutants, especially for non-Gaussian distributed health outcomes, e.g. counts of mortality or hospital admissions.

5.1.1.2 Review of Spatiotemporal Regression Models Spatiotemporal regression models have been developed for environmental and socioeconomic epidemiology, ecology and agriculture. [Zhu et al., 1999, Xia and Carlin, 1998, Wikle, 2003, Cressie and Majure, 1997]. However, in these models, spatiotemporal autocorrelation leads to a complicated variance-covariance matrix for the dependent variable in a regression model. Therefore, a spatiotemporal regression usually involves a latent process (or random effect) with a covariance function parameterized by temporal and spatial coordinates [Ma, 2003] or a hierarchy of nested temporal and spatial processes [Waller et al., 1997]. As estimating covariance functions requires inverting large covariance matrices, model inference for a spatiotemporal regression is usually performed using approximating methods [Genton, 2007, Sang and Huang, 2012] or Bayesian methods, e.g. Monte Carlo Markov Chain [Cressie and Wikle, 2011]. However, for most studies, full inference for hidden processes is not necessary, as the estimating regression coefficients is usually the major aims in practical respect. Therefore, this paper proposes to develop marginal estimators for the coefficients in a Poisson regression model of spatiotemporal counts, without specific inference for the variance-covariance matrix.

Even though marginal estimation for spatial or spatiotemporal regression coefficients has rarely been explored, mature methods have been have developed for the time-series regression model. Zeger et al. applied generalized estimating equations to a parameter-driven Poisson model of time-series counts using a working correlation matrix of an autoregressive filter to control temporal autocorrelation [Zeger, 1988]. In this section, we are going to extend Zeger's methods from time-series Poisson regression to spatiotemporal Poisson regression. Instead of an autoregressive filter, we introduce a working correlation matrix of a structural vector autoregressive (SVAR) filter to control spatiotemporal dependences in the Poisson counts. Thus, we name our method as spatiotemporal generalized estimating equations, which avoids inverting a large variance-covariance matrix in construct to traditional spatiotemporal models.

5.1.2 Data Description

The statistical method in this section is motivated by a study to associate the six air pollutants ($PM_{2.5}$, PM_{10},O_3, NO_2, SO_2, CO) to daily mortality counts of 213 USPS ZIP Code Tabulation Areas (ZCTA) in the Pittsburgh region, from 1999 to 2008. As the mortality counts are highly correlated both in spatial and temporal dimensions, we had to take spatiotemporal autocorrelation into consideration.

5.1.2.1 Mortality Data The 1999-2008 mortality records were collected from the Pennsylvania Department of Health for a study domain of three counties, Allegheny, Washington and Westmoreland as shown in Figure 1. The major cause of death was coded using ICD-10 and death address was coded by the USPS ZIP code. We excluded the two areas with no contiguity (zip codes: 15618 and 15690). Finally, we aggregated the 217,719 death records into daily counts for the 213 ZCTA and 4018 days. The time series of daily aggregated mortality counts over the study domain is displayed in Figure 2.

5.1.2.2 Demographic Data In order to adjust the demographic information for mortality risks, we calculated the expected daily mortality for each ZCTA. We collected population sizes by sex and age groups for each ZCTA from two 2010 census data (https://www.census.gov/). Assuming that the population was stable over time in our study domain, in order to compare mortality risks for various diseases between ZCTAs, we generated standardized mortality ratios (SMRs) for each day and each ZCTA as

$$\text{SMR}_{s,t} = y_{s,t}/e_s,$$

where (s, t) are regular spatial and temporal indexes respectively and e_s is the expected death count in ZCTA s, calculated based on demographic data for sex and age analogously to SIR (Equation 2.13) in Section 2.2.3. In Poisson regression, we are going to use e_s as a offset for ZCTA s to control population size adjusted by sex and age and interpret the regression coefficient as an increase of log(SMR) for per unit increase in the covariate. The spatial patterns of SMRs for total mortality, cardiovascular diseases, respiratory diseases and

cancer are shown in Figure 26. In the following analysis, we excluded the isolated ZCTAs (islands in SMR maps).

5.1.2.3 Environmental Data The environmental data of daily air pollutants and temperature (collected from the NOAA Global Historical Climatology Network) were generated by averaging the $1km \times 1km$ spatiotemporal predictions in Section 2.1.4 by ZCTAs. The spatial patterns of ZCTA level air pollutants are shown in Figure 27. The uncertainty of predictions of various ZCTAs depends on their locations relative to the air monitors, therefore exposure assessments may be more accurate in the city of Pittsburgh (yellow and blue polygons in Figure 1) where air monitors are clustered than Westmoreland and Washington counties, where there are few monitors. Considering possibility of spatially heterogeneous exposure misclassification, we will restrict our analysis to three different areas: the city of Pittsburgh, Allegheny County and the three county study domain as shown in Figure 1.

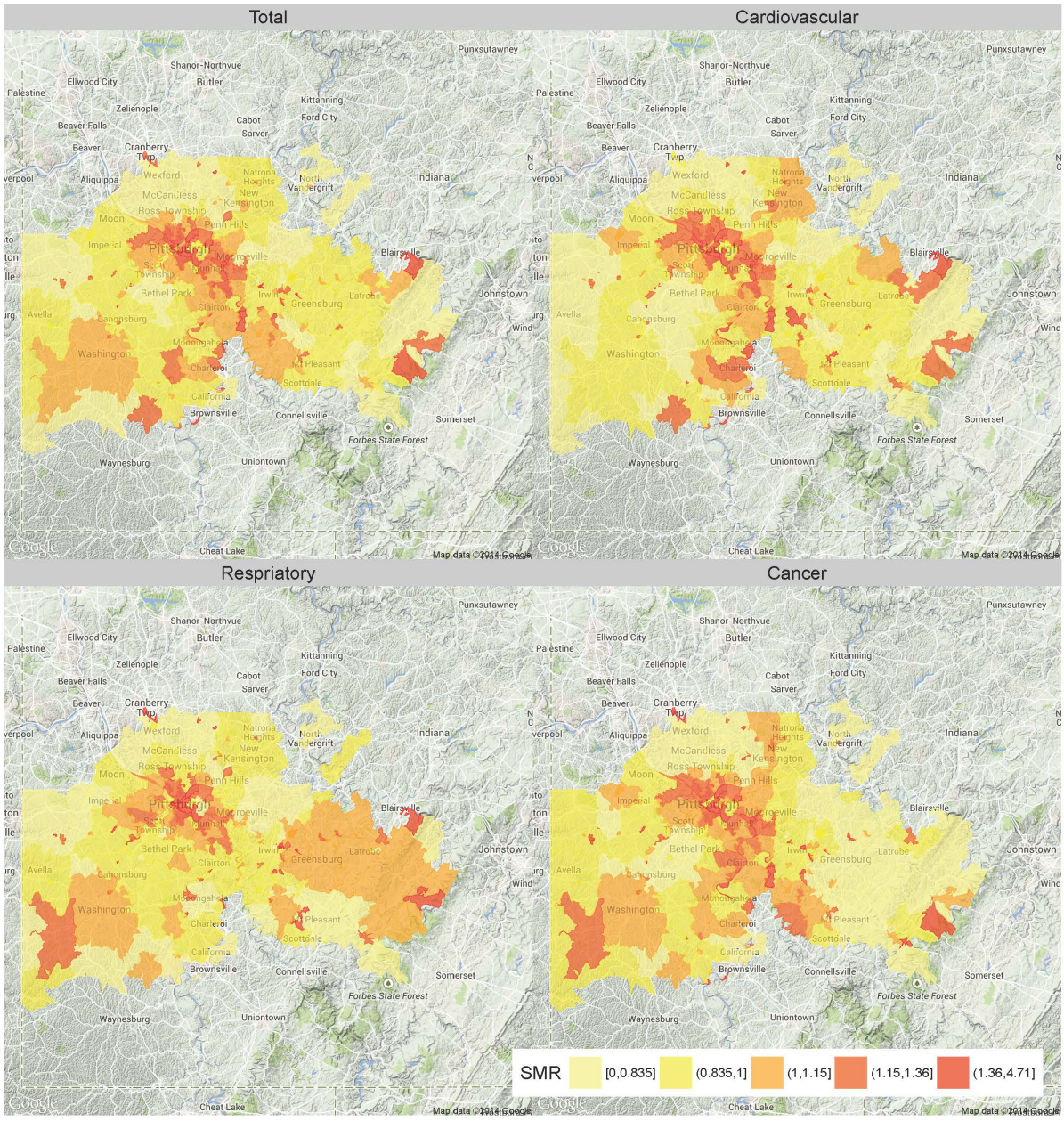
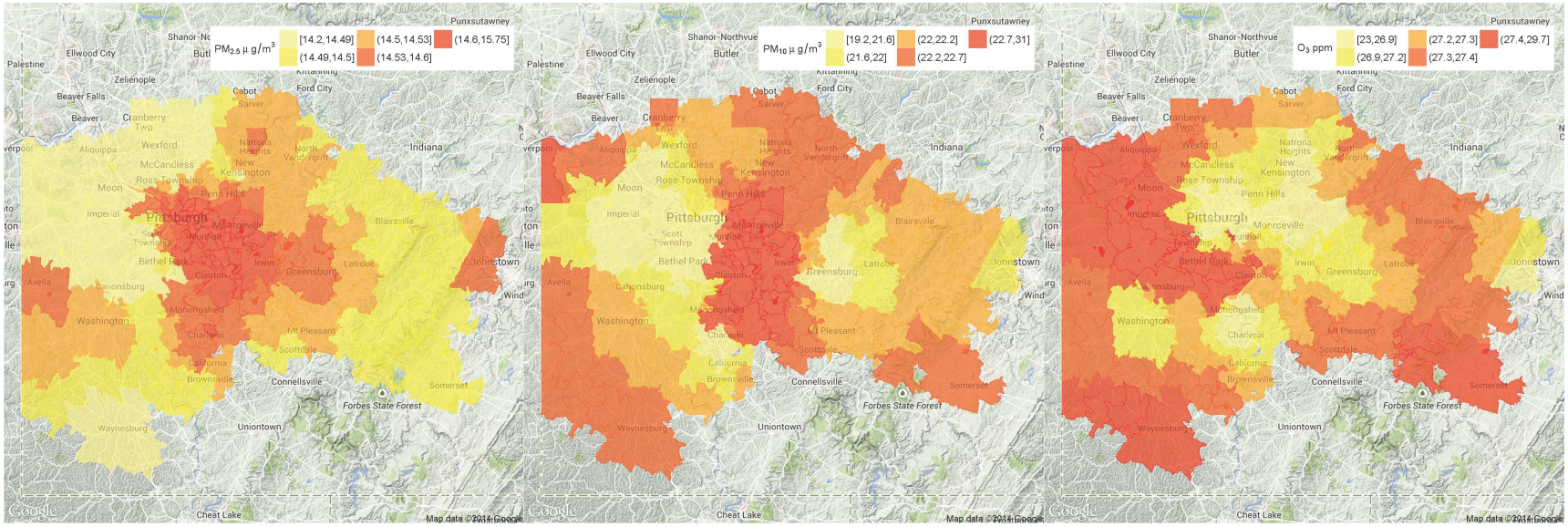
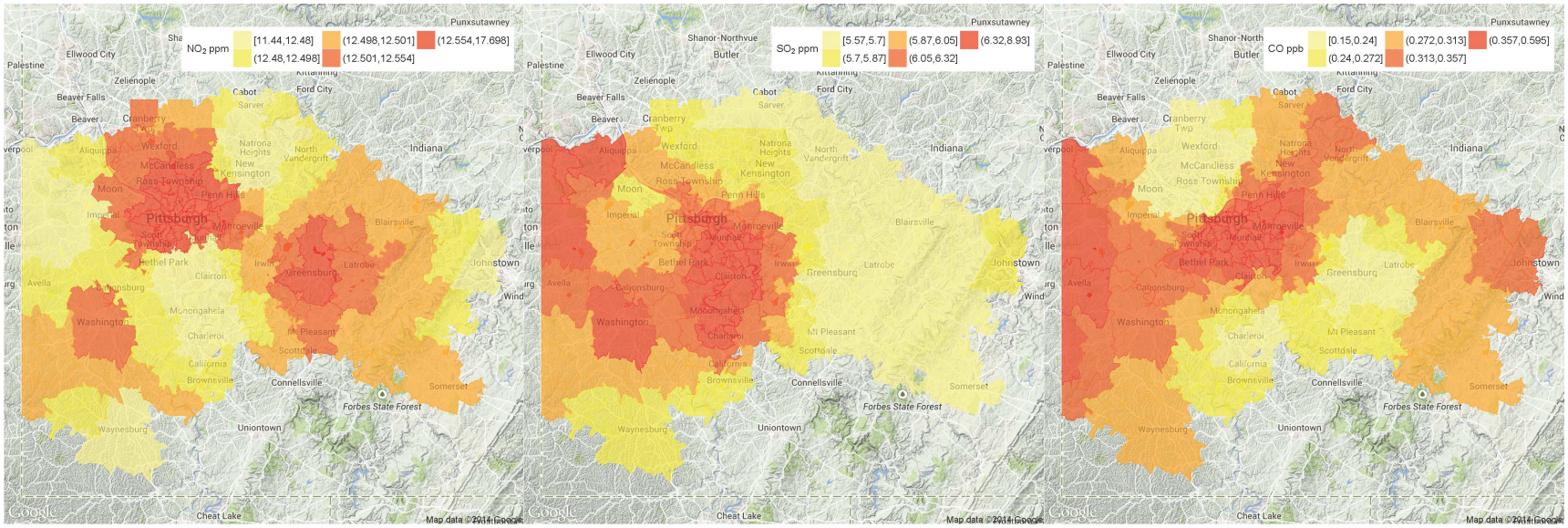


Figure 26: Spatial patterns of averaged sex and aged adjusted standardized mortality ratios (SMRs) for total mortality, cardiovascular diseases, respiratory diseases and cancer for all ZCTAs in the Pittsburgh region area from 1999 to 2008.



(a) $PM_{2.5}$ (b) PM_{10} (c) O_3



(d) NO_2 (e) SO_2 (f) CO

Figure 27: Spatial patterns of air pollutants where the average level is shown for each ZCTA.

5.2 STATISTICAL MODEL: SPATIOTEMPORAL GENERALIZED ESTIMATING EQUATIONS

5.2.1 Model Assumptions and Specification

In this section, we will describe a regression model to associate pollutant exposure to daily counts of mortality in each ZCTA by adding a latent spatiotemporal process, $\nu_{s,t}$ into a Poisson generalized linear model (GLM):

$$\mathrm{E}(y_{s,t}|\nu_{s,t}) = \exp(x'_{s,t}\beta)\nu_{s,t} = \mu_{s,t}\nu_{s,t}, \quad \mathrm{var}(y_{s,t}|\nu_{s,t}) = \mu_{s,t}\nu_{s,t}; \tag{5.1}$$

where (s,t) is the spatiotemporal coordinate for mortality counts $(y_{s,t})$ and spatiotemporal covariates $(x_{s,t})$. Assume that $\nu_{s,t}$ is an unobserved stationary process with $\mathrm{E}(\nu_{s,t}) = 1$ and with a two-dimensional covariance function $\mathrm{cov}(\nu_{s,t}, \nu_{s+\Delta_s,t+\Delta_t}) = \sigma^2\rho(\Delta_s, \Delta_t; \theta)$. Thus the marginal expectation, variance and correlation function of $y_{s,t}$ can be derived as following:

$$\mathrm{E}(y_{s,t}) = \exp(x'_{s,t}\beta) = \mu_{s,t} \quad \mathrm{var}(y_{s,t}) = \mu_{s,t} + \sigma^2\mu_{s,t}^2 \tag{5.2}$$

$$\mathrm{corr}(y_{s,t}, y_{s+\Delta_s,t+\Delta_t}) = \frac{\rho(\Delta_s, \Delta_t)}{\{[1 + (\sigma^2\mu_{s,t})^{-1}][1 + (\sigma^2\mu_{s+\Delta_s,t+\Delta_t})^{-1}]\}^{\frac{1}{2}}}. \tag{5.3}$$

In this model, the latent spatiotemporal process $\nu_{s,t}$ account for both over-dispersion and autocorrelation of the count data $(y_{s,t})$, but the a stationary latent process $\nu_{s,t}$ do not lead to a stationary spatiotemporal autocorrelation in $y_{s,t}$. In addition, the restriction $\mathrm{E}(\nu_{s,t}) = 1$ guarantees interpretable coefficients (β) for the regression model: the change of health outcomes $(y_{s,t})$ is proportional to the exponentiation-scale of coefficients but not depends on the latent process $(\nu_{s,t})$.

The above model is desirable for epidemiological interpretation but usually requires considerable computational efforts to infer the latent process, especially for our massive spatiotemporal dataset. However, for epidemiological purposes, we only need to estimate the coefficient parameters (β) and their confidence intervals by treating the parameters (θ, σ^2) in the latent process $\nu_{s,t}$ as a nuisance. Therefore, we developed estimating equations for the coefficients (β) analogous to Zeger's method [Zeger, 1988].

5.2.2 Model Inference

5.2.2.1 Generalized Estimating Equations for Spatiotemporal Poisson Counts

For the m areal locations and n temporal points, letting

$$\boldsymbol{y} = (\underbrace{y_{s_1,t_1}, \cdots, y_{s_m,t_1}}_{m}, \cdots, \cdots, \underbrace{y_{s_1,t_n}, \cdots, y_{s_m,t_n}}_{m})';$$

$$\boldsymbol{X} = (\underbrace{\boldsymbol{x}_{s_1,t_1}, \cdots, \boldsymbol{x}_{s_m,t_1}}_{m}, \cdots, \cdots, \underbrace{\boldsymbol{x}_{s_1,t_n}, \cdots, \boldsymbol{x}_{s_m,t_n}}_{m})';$$

$$\boldsymbol{\mu} = (\underbrace{\mu_{s_1,t_1}, \cdots, \mu_{s_m,t_1}}_{m}, \cdots, \cdots, \underbrace{\mu_{s_1,t_n}, \cdots, \mu_{s_m,t_n}}_{m})', \text{ where } \mu_{s,t} = \exp(\boldsymbol{x}_{s,t}\boldsymbol{\beta});$$

$$\boldsymbol{V} = \text{var}(\boldsymbol{y}) = \boldsymbol{C} + \sigma^2 \boldsymbol{C}\boldsymbol{R}_{\boldsymbol{\nu}}\boldsymbol{C}, \text{ where } \boldsymbol{C} = \text{diag}(\boldsymbol{\mu}), \ \boldsymbol{R}_{\boldsymbol{\nu}} \text{ is determined on } \rho(\cdot;\theta);$$

the quasi-likelihood estimating equation can be constructed as following:

$$\boldsymbol{\psi}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}} \boldsymbol{V}^{-1}(\boldsymbol{\beta}, \theta, \sigma^2)(\boldsymbol{y} - \boldsymbol{\mu}) = 0. \tag{5.4}$$

The variance-covariance of $\boldsymbol{y}$ is parameterized by regression coefficients ($\boldsymbol{\beta}$) and nuisance parameters (σ^2, θ), which requires an iterative weighted process to solve the estimating equations. However, computational complexity of inverting the $(mn \times mn)$ matrix $\boldsymbol{V}$ is massive. Thus we consider an approximation to $\boldsymbol{V}$ using "working correlation matrix"approach: $\boldsymbol{V} \approx \boldsymbol{V}_R = \boldsymbol{D}^{\frac{1}{2}}(\boldsymbol{\beta})\boldsymbol{R}(\boldsymbol{\alpha})\boldsymbol{D}^{\frac{1}{2}}(\boldsymbol{\beta})$, where $\boldsymbol{D}$ is a diagonal matrix with marginal variance of $\boldsymbol{y}$ and $(\boldsymbol{D} = \boldsymbol{C} + \sigma^2\boldsymbol{C}\boldsymbol{C})$. The working correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$ provides a guess of the true autocorrelation of $\boldsymbol{y}$. The $\boldsymbol{A}(\boldsymbol{\beta})$ and $\boldsymbol{B}(\boldsymbol{\beta})$ components for the M-estimators can be derived as following (see chapter 7 of [Boos and Stefanski, 2013] for details of M-estimators):

$$\boldsymbol{A}(\boldsymbol{\beta}) = \lim_{n\to\infty} \left(\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_R^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \Big/ n \right), \quad \boldsymbol{B}(\boldsymbol{\beta}) = \lim_{n\to\infty} \left(\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_R^{-1} \boldsymbol{V} \boldsymbol{V}_R^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \Big/ n \right).$$

Let $\boldsymbol{\beta}$ be the solution of the estimating equations (Equation 5.4). Therefore we can conclude that under mild regularity conditions, the existence of the limits in $[\boldsymbol{A}(\boldsymbol{\beta}), \boldsymbol{B}(\boldsymbol{\beta})]$ and given that $(\hat{\theta}, \hat{\sigma^2})$ are $\sqrt{n}$-consistent estimators of the nuisance parameters, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically multivariate Gaussian with zero mean and covariance matrix $\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\boldsymbol{A}(\boldsymbol{\beta})^{-1}$. The proof is similar to [Zeger, 1988] and is omitted here.

5.2.2.2 Spatiotemporal Working Correlation Matrix $R(\alpha)$ Assumption: Vector Autoregressive Process In [Zeger, 1988], the nonstationary autocorrelation Poisson process was approximated by a stationary p-ordered autoregressive process: $\boldsymbol{V}_R^{-1} \approx \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}'\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$, where $\boldsymbol{L}$ is the matrix form for an autoregressive filter, "$y_t - \alpha_1 y_{t-1} -, \cdots, -\alpha_p y_{t-p} \quad (t > p)$". While, we approximate the nonstationary spatiotemporal autocorrelation by a stationary vector autoregressive (VAR) process [Lütkepohl, 2006, Johansen, 1995], which has been popular in both analyzing multivariate time series in economics [Sims, 1980] since 1980s and has been applied to capture the spatiotemporal dependence [Di Giacinto, 2010].

Redefine an m-dimensional vector for all the areal observations at time t as follows:

$$\boldsymbol{y}_t = (y_{s_1,t}, \cdots, y_{s_m,t})', \text{ thus } \boldsymbol{y} = (\boldsymbol{y}'_{t_1}, \cdots, \boldsymbol{y}'_{t_n})'.$$

Therefore, a VAR(p) process can be described as:

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{H}_1\boldsymbol{y}_{t-1} +, \cdots, +\boldsymbol{H}_p\boldsymbol{y}_{t-p} + \boldsymbol{\eta}_t; \;\; E(\boldsymbol{\eta}_t) = 0, \; E(\boldsymbol{\eta}_t\boldsymbol{\eta}'_t) = \boldsymbol{\Sigma}_{\boldsymbol{\eta}}, \; E(\boldsymbol{\eta}_t\boldsymbol{\eta}'_{\boldsymbol{s}}) = 0. \quad (5.5)$$

In a VAR(p) process, the $\boldsymbol{H}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ capture the temporal and spatial autocorrelations. However, if we use the $\boldsymbol{H}$ matrices to construct a VAR(p) filter, the filtered variance-covariance matrix will be a block matrix, which is not the most efficient approach. Thus, we considered a structural vector autoregressive (SVAR) model [Kilian, 2011]:

$$\boldsymbol{F}\boldsymbol{y}_t = \boldsymbol{c}^* + \boldsymbol{H}_1^*\boldsymbol{y}_{t-1} +, \cdots, +\boldsymbol{H}_p^*\boldsymbol{y}_{t-p} + \boldsymbol{F}\boldsymbol{\eta}_t; \;\; \boldsymbol{H}^* = \boldsymbol{F}\boldsymbol{H}, \; E(\boldsymbol{F}\boldsymbol{\eta}_t\boldsymbol{\eta}'_t\boldsymbol{F}') = \boldsymbol{F}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{F}'. \quad (5.6)$$

In Equation 5.6, if we choose a specific form of the matrix $\boldsymbol{F}$, for example, the Cholesky decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, the $E(\boldsymbol{F}\boldsymbol{\eta}_t\boldsymbol{\eta}'_t\boldsymbol{F}')$ can be transformed as identify matrix, $\boldsymbol{I}_{\boldsymbol{m}}$. Let the matrix $\boldsymbol{L}$ to define a SVAR(p) filter: $\boldsymbol{F}\boldsymbol{y}_t - \boldsymbol{H}_1^*\boldsymbol{y}_{t-1} -, \cdots, -\boldsymbol{H}_p^*\boldsymbol{y}_{t-p}$, thus

$$\boldsymbol{V}_R^{-1} \approx \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}'\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}.$$

5.2.2.3 Iterative Weighted Least-squares Methods to Solve Estimating Equations

Setting a working correlation matrix using an SVAR(p) filter, we can solve the spatiotemporal estimating equations through an iterative weighted least-squares procedure:

$$\hat{\boldsymbol{\beta}}^{(j+1)} = \left\{ \left(\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)' \left(\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right) \right\}^{-1} \left(\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right)' (\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Z}), \tag{5.7}$$

$$\boldsymbol{Z} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}} \boldsymbol{\beta} + (\boldsymbol{y} - \boldsymbol{\mu}).$$

At the right side of Equation 5.7, we use the coefficients that are estimated in the last iteration, $\hat{\boldsymbol{\beta}}^{(j)}$. Thus this algorithm involves the following steps:

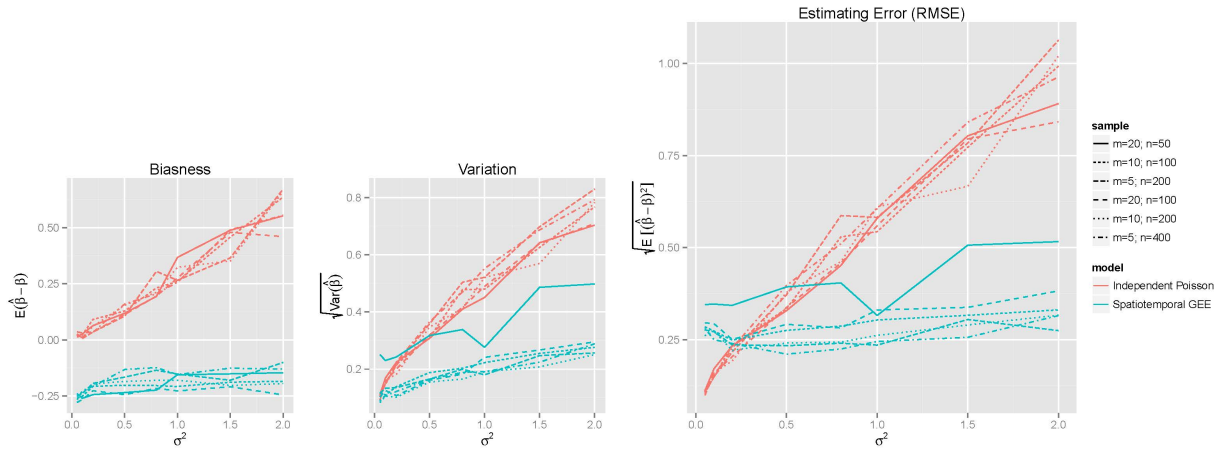
1. Weight the current values of $\partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ and $\boldsymbol{Z}$ by $\boldsymbol{D}^{-\frac{1}{2}}$, which is given by

$$\boldsymbol{D}^{-\frac{1}{2}} = \operatorname{diag}\left(1/\sqrt{\mu_{s_1,t_1} + \sigma^2 \mu_{s_1,t_1}^2}, \cdots, 1/\sqrt{\mu_{s_m,t_n} + \sigma^2 \mu_{s_m,t_n}^2}\right),$$

$$\mu_{s,t} := \hat{\mu}_{s,t}^{(j)} = \exp(\boldsymbol{x}_{s,t}\hat{\boldsymbol{\beta}}^{(j)}), \quad \sigma^2 := \widehat{\sigma^2}^{(j)} = \sum_{\substack{s=1 \\ t=1}}^{mn} \left\{ [y_{s,t} - \hat{\mu}_{s,t}^{(j)}]^2 - \hat{\mu}_{s,t}^{(j)} \right\} \Big/ \sum_{\substack{s=1 \\ t=1}}^{mn} [\hat{\mu}_{s,t}^{(j)}]^2;$$

2. Estimate the SVAR(p) filter matrix $\boldsymbol{L}$ based on the current standardized residuals using existing methods, e.g. the method in [Pfaff, 2008] provided by the R package `vars` [Bernhard Pfaff, 2008];
3. Apply the filter to $\boldsymbol{D}^{-\frac{1}{2}}\partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ and $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Z}$;
4. Solve the least squares equations;
5. Iterate the above steps to convergence.

In our following analysis, we focus on an SVAR(1) filter, which may be more useful in applications. We do not discuss further the computational complexity and convergence property of the algorithm in the present paper. However, we found that the algorithm converges in less than 10 steps. Estimating the SVAR(p) filters ($\boldsymbol{L}$) is the most computationally expensive part of our algorithm, which limits our algorithm to the case where size of temporal observations (n) is larger than that of spatial observations (m). In addition, due to the robust estimator (or switch estimator), $\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\boldsymbol{A}(\boldsymbol{\beta})^{-1}$ is not appropriate for a massive variance-covariance $\boldsymbol{V}$ in spatiotemporal analysis. Thus, we use $\boldsymbol{A}(\beta)^{-1}/n$ to estimate the standard errors of β.



(a) RMSE

(b) Variance & Biasness

Figure 28: A simulation example to compare estimating accuracy between spatiotemporal generalized estimating equations and independent Poisson regression.

5.3 A SIMPLE SIMULATION

To illustrate the performances of the spatiotemporal generalized estimating equations, we generated a simple simulation study to compare this model with the independent Poisson regression (fitted by `glm` in R). The efficiency of our estimators was influenced by two factors: the sample size (including the total sample size mn and the sample size ratio of spatial to temporal observations m/n) and the relative magnitude of spatiotemporal noise (relative magnitude of spatiotemporal random effects $\nu_{s,t}$ to fixed effects $X_{s,t}\beta$). We simulate x_{st} and $\boldsymbol{\nu} = (\nu_{s_1,t_1}, \cdots, \nu_{s_m,t_n})'$ as follows:

$$x_{st} = \sin\left(\frac{2\pi t}{\max(t)}\right) + \frac{\omega}{5}, \text{ where } \omega \sim N(0,1),$$

$$\boldsymbol{\nu} \sim N_n(0, \boldsymbol{\Sigma}_\nu), \text{ where } (\boldsymbol{\Sigma}_\nu)_{i,j} = \sigma^2 \rho(|s_i - s_j|_2, |t_i - t_j|_2) \text{ and}$$

$$\rho(\Delta_s, \Delta_t) = k_s \exp(-\frac{\Delta_s}{h_s}) + k_t \exp(-\frac{\Delta_t}{h_t}) + k_{st} \exp(-\frac{\Delta_s}{h_s}) \exp(-\frac{\Delta_t}{h_t}),$$

in which, we mimic variations in air pollutants using a $\sin(\cdot)$ function plus a small normal noise and introduce a stationary spatiotemporal latent process using a product-sum correlation function [De Iaco et al., 2002b]. In order to limit flexibility in covariance function, we fixed correlation parameters: $h_s = 0.3\max(\Delta_s), h_t = 0.3\max(\Delta_t), k_s = k_t = 0.45$ and $k_{st} = 0.1$ and used only the parameter σ^2 to control the magnitude of spatiotemporal noise. Therefore, in the simulation example, we mainly explored the relationship between estimation accuracy (including biasness $\mathrm{E}(\hat{\beta} - \beta)$, variance $\mathrm{Var}(\hat{\beta})$ and errors $\mathrm{E}[(\hat{\beta} - \beta)^2]$) and three parameters including m, n and σ^2, as displayed in Figure 28. According to the simulation results, we can conclude that: (1) under the case of large spatiotemporal noise, spatiotemporal generalized estimating equations approach is better than independent Poisson regression due to increased biasness and variance of the latter model with increasing spatiotemporal variance (σ^2)l, (2) spatiotemporal generalized estimating equations approach usually underestimates the coefficients, and (3) the estimation errors of our model decrease with a larger ratio n/m and larger sample size mn.

5.4 RESULTS

5.4.1 Descriptive Analysis

The summary statistics for mortality and air pollutants are shown in Table 10. Temporal autocorrelations are described by the auto-correlation function (ACF) and partial autocorrelation function (Partial ACF) of the aggregated counts of mortality within our study domain, as shown in Figure 29. The ACF and Partial ACF plots show that mortality counts are highly autocorrelated in the temporal dimension. Spatial autocorrelation is described by the statistics Moran's I and Geary's C [Cliff and Ord, 1981], for averaged SMRs for each ZCTA as shown in Table 11. I ranges from -1 to 1 and $I > 0$ indicates positive spatial autocorrelation (which means similar values are spatially clustered together); C ranges from 0 to 3 and $C < 1$ indicates positive spatial autocorrelation. The Monte Carlo permutation tests showed that all the spatial autocorrelations of SMRs were statistically significantly

positive at the significance level of 0.1. The above analysis suggests that the mortality data are highly correlated both in spatial and temporal dimensions, which should not be ignored in a regression model.

5.4.2 Regression Results

In the Poisson regression models, we involve three covariates: (1) one type of air pollutant, (2) temperature and (3) a smoothed term for a long term trend (natural spline with 50 degrees of freedom) and the offset for expected mortality. We compare the estimated coefficients of air pollutants between spatiotemporal generalized estimating equations and independent Poisson regression as shown in Figure 30. For the particulate matter ($PM_{2.5}$ and PM_{10}), the results of the independent Poisson regression and our spatiotemporal generalized estimating equations are consistent with each other, while for the gaseous pollutants, the regression results diverge from each other. Based on the regression results, we find that estimators of the spatiotemporal generalized estimating equations are usually lower than those of the independent Poisson models except for O_3. For O_3, independent Poisson results are statistically significantly negative, which indicates a protective effect of O_3 and is opposite to common sense and previous results [Ito et al., 2005]; while, spatiotemporal generalized estimating equations report near-zero health effects. In addition, the estimated coefficients are more consistent among the three areas' analysis in results of spatiotemporal generalized estimating equations than the independent Poisson models. However, the spatiotemporal generalized estimating equations reported fewer statistically significant associations than did the independent Poisson regression, possibility due to the biased estimation from the method as shown in the simulation study.

5.4.3 Lag Analysis

As a lag period of a few days may exist between the exposure to air pollutants and its acute effects on mortalities, we also associated SMRs with lagged air pollutants from one to seven-days lag as shown in Figure 31. In order to save computing time and minimize potential exposure misclassification, we did the analysis in this section using only Allegheny

Table 10: Summary statistics for air pollutants and mortalities.

Variable	Area	Total		Jan.-Mar.		Apr.-Jun.		Jul.-Sep.		Oct.-Dec.	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
$PM_{2.5}$ $\mu g/m^2$	Study domain	14.58	7.97	12.37	5.51	14.7	8.30	18.92	9.31	12.27	6.22
	Allegheny	14.66	8.04	12.45	5.58	14.76	8.38	18.99	9.38	12.38	6.31
	Pittsburgh	14.67	8.05	12.48	5.60	14.74	8.38	19.02	9.42	12.41	6.30
PM_{10} $\mu g/m^2$	Study domain	22.42	12.14	18.36	9.52	24.10	12.57	28.17	12.61	19.01	10.83
	Allegheny	22.59	12.99	18.41	10.29	24.27	13.43	28.33	13.41	19.30	11.99
	Pittsburgh	21.27	11.56	17.57	9.04	22.60	11.73	26.81	12.26	18.05	10.38
O_3 *ppb*	Study domain	26.63	9.99	23.23	7.38	34.69	7.75	30.85	8.45	17.78	6.32
	Allegheny	26.26	10.13	22.72	7.52	34.3	7.90	30.69	8.57	17.34	6.42
	Pittsburgh	24.50	10.37	20.66	7.61	32.17	8.21	29.67	9.01	15.50	6.47
NO_2 *ppb*	Study domain	13.11	5.24	15.70	5.83	11.50	4.23	10.66	3.44	14.59	5.37
	Allegheny	13.52	5.54	16.11	6.11	11.94	4.63	11.08	3.85	14.98	5.66
	Pittsburgh	15.60	6.48	18.13	7.01	14.23	5.87	13.19	5.05	16.88	6.58
SO_2 *ppb*	Study domain	6.64	3.81	8.08	4.32	5.58	2.77	5.48	2.72	7.44	4.37
	Allegheny	6.82	4.18	8.22	4.69	5.67	3.15	5.60	3.04	7.81	4.81
	Pittsburgh	6.82	4.18	8.14	4.84	5.65	3.11	5.74	3.11	7.77	4.70
CO *ppm*	Study domain	0.29	0.37	0.34	0.39	0.26	0.36	0.25	0.32	0.31	0.40
	Allegheny	0.34	0.48	0.39	0.50	0.30	0.47	0.30	0.41	0.38	0.53
	Pittsburgh	0.42	0.53	0.48	0.57	0.37	0.50	0.36	0.43	0.47	0.59
Temperature °C	Study domain	7.87	9.04	-1.32	6.29	11.93	5.63	17.02	3.71	3.72	6.34
	Allegheny	8.01	9.04	-1.25	6.16	12.09	5.60	17.23	3.66	3.84	6.27
	Pittsburgh	8.10	9.03	-1.16	6.14	12.18	5.60	17.33	3.63	3.92	6.27
Daily SMR	Study domain	1.12	7.37	1.25	7.76	1.10	7.32	1.03	7.05	1.12	7.32
	Allegheny	1.14	5.15	1.26	5.22	1.11	5.06	1.04	4.84	1.14	5.46
	Pittsburgh	1.31	2.31	1.44	2.39	1.30	2.35	1.21	2.19	1.30	2.30
Daily SMR of	Study domain	1.13	12.12	1.26	12.45	1.12	12.62	1.02	11.52	1.12	11.88
circulatory diseases	Allegheny	1.15	8.61	1.32	9.11	1.10	7.84	1.04	8.36	1.14	9.07
	Pittsburgh	1.33	3.92	1.49	4.13	1.32	4.05	1.17	3.63	1.33	3.84
Daily SMR of	Study domain	1.13	15.12	1.20	15.51	1.12	14.93	1.11	14.64	1.12	15.37
cancer	Allegheny	1.19	11.53	1.20	10.20	1.24	14.09	1.13	9.54	1.19	11.76
	Pittsburgh	1.31	4.49	1.36	4.55	1.27	4.40	1.29	4.46	1.30	4.55
Daily SMR of	Study domain	1.10	24.40	1.36	26.65	1.09	24.43	0.87	21.28	1.08	24.98
respiratory diseases	Allegheny	1.09	15.58	1.26	10.77	1.09	16.26	0.87	12.20	1.16	20.97
	Pittsburgh	1.28	7.45	1.59	8.19	1.24	7.55	1.08	6.98	1.21	7.01

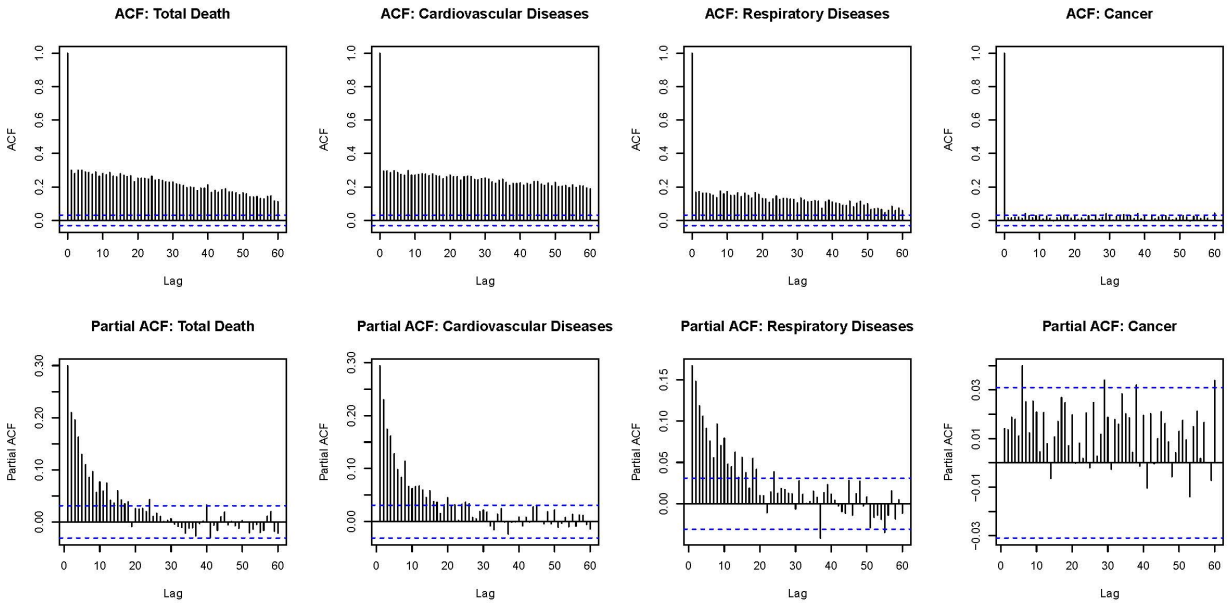
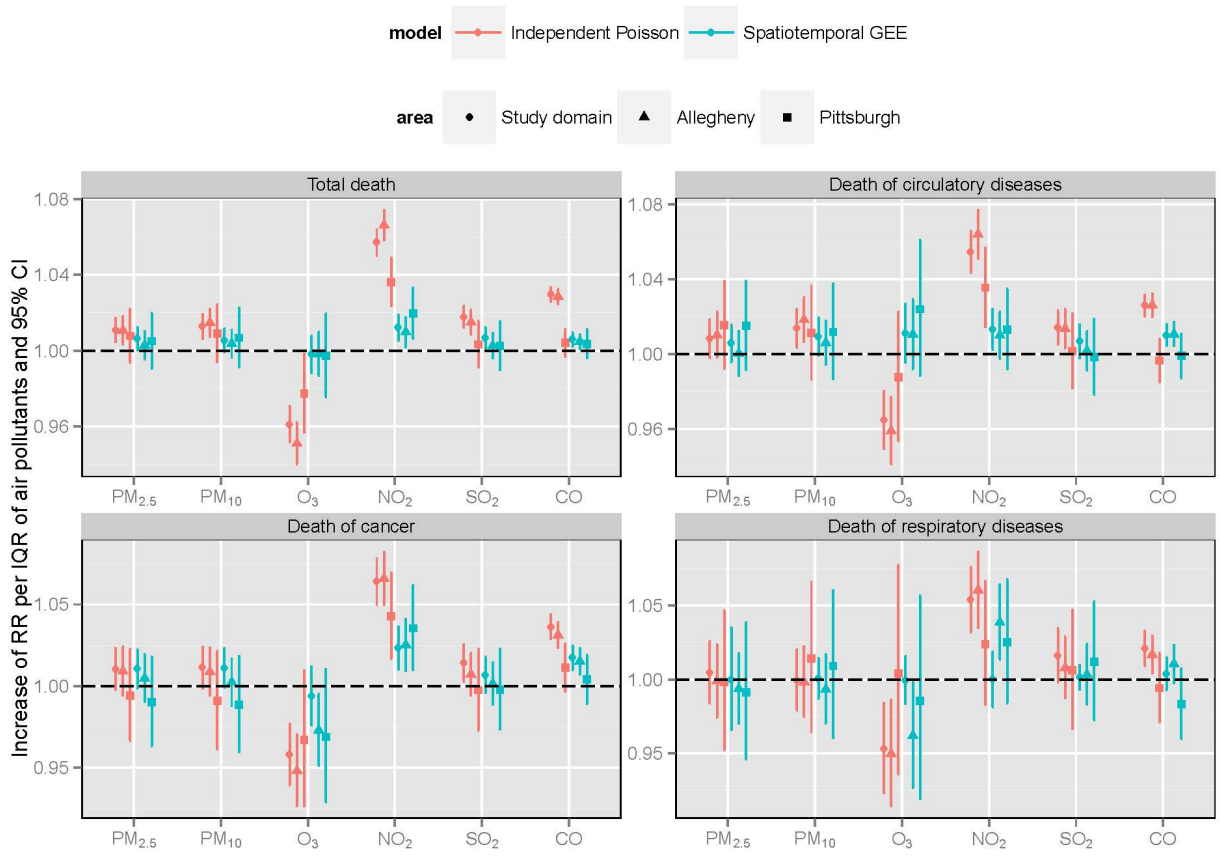


Figure 29: Auto-correlation functions and partial auto-correlation functions for daily aggregated mortality counts over study domain.

Table 11: Statistics and p-values of Monte Carlo permutation tests for positive spatial autocorrelation.

SMR	Moran's I		Geary's C	
	I	p-value	C	p-value
Total death	0.1041	0.0118	0.8613	0.0736
Circulatory diseases	0.0868	0.0292	0.8572	0.0414
Cancer	0.0861	0.0272	0.8327	0.0348
Respiratory diseases	0.0284	0.0648	0.5759	0.0088

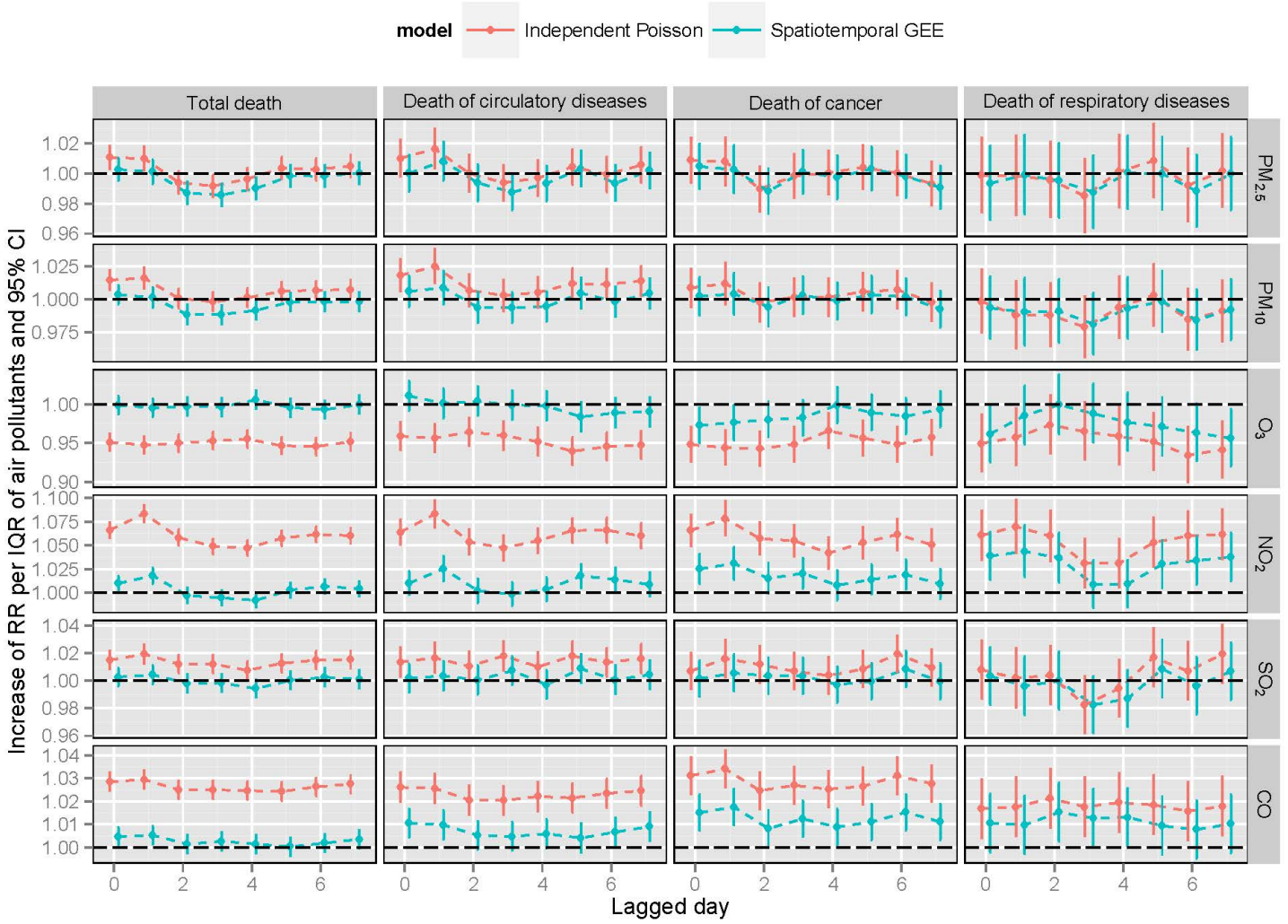
For Moran's I, the alternative hypothesis is $I > 0$; for Geary's C, the alternative hypothesis is $C < 1$. The analysis is performed using R-package `spdep` [Roger Bivand, 2014].



We present results for three areas by shapes and two types of models by colors.

Figure 30: Increase of relative risk for mortality per IQR due to air pollutants and their corresponding 95% confidence intervals.

County's data. The lagged patterns of total mortality, mortality of cardiovascular diseases and mortality of cancer are consistent with each other and all pollutants' peak effects on the two health outcomes appear within one day's lag except for O_3. For mortality due to respiratory diseases, the lagged patterns diverge from each other and only NO_2 and CO are statistically significantly associated with respiratory diseases and their peak effects are reported at lag of 1 and 2 days.



We use Allegheny County's data in the lagged period analysis and compare the independent Poisson model with the spatiotemporal GEE method.

Figure 31: Increase of relative risk of mortalities per IQR for lagged 0-7 days air pollutants and their 95% confidence intervals.

5.5 DISCUSSION

5.5.1 Limitation of Spatiotemporal Generalized Estimating Equations

Based on our simulation results, our spatiotemporal generalized estimating equations approach is limited because of the potential underestimation of the regression of coefficients. The overfit of spatiotemporal structure (VAR process) of the residuals may contribute to the shrinking of our regression coefficients to zero. When estimating the spatiotemporal autocorrelation through fitting the auto-regression coefficients of a VAR process, we do not make any structural assumption about the spatiotemporal covariance matrix, which is unusual in spatiotemporal analysis. For example, in a first order CAR model, spatial dependences are assumed to exist between neighboring areas and all the other areas are assumed to be spatially independent. Introducing the sparsity of spatial dependence, we can add a hard-threshold when optimizing Equation 5.6 to force some of the elements in matrices $\boldsymbol{F}$ and $\boldsymbol{H}$ to be zero according to the spatial neighbor structure. In order to avoid manually selecting the spatiotemporal dependence structure, we can also apply a sparse vector autoregressive model [Davis et al., 2012] instead of a regular autoregressive model.

Another disadvantage is that our statistical method is only appropriate for the case where the size of the temporal observations (n) is larger than that of spatial observations (m). This restriction is also caused by fitting a VAR filter and intended to guarantee a unique solution for matrices $\boldsymbol{F}$ and $\boldsymbol{H}$. Therefore, the structural or sparse spatiotemporal dependence assumption may also help to improve the model performance in the case where n and m are comparable.

5.5.2 Limitation of the Study Design

The major weakness of the study design is due to the geocoding. To protect the confidentiality of subjects in our study, exact home addresses are not available, but only the USPS ZIP code. Therefore, we had to geocode the mortality records using 2010 ZCTA maps. However, on the one hand, USPS ZIP code may have varied during our study period, which can lead to potential exposure misclassification. On the other had, comparing with other

areal units such as census tracts, ZCTA is disadvantageous in two aspects: (1) ZCTAs are irregular shapes and in some of the tiny ZCTAs, their SMRs are possible outliers because of extremely small number of residences; (2) fewer socioeconomic factors were available at the ZIP code level and therefore we may fail to adjust for some confounding effects in our study.

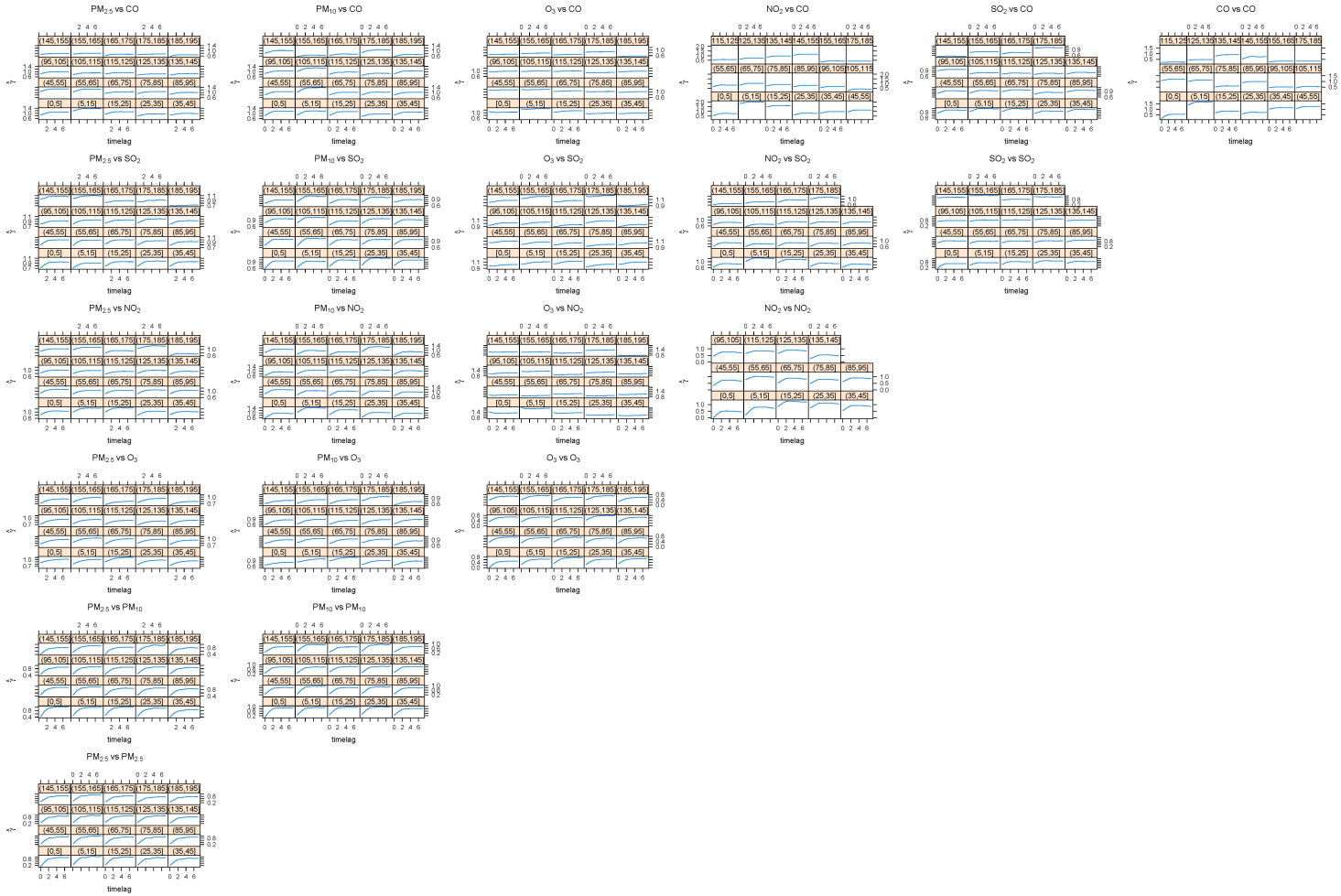
In addition, through calculating expected mortalities in each ZCTA, we adjusted for sex and age as potential confounding effects to the mortality risks, but ignored other demographic confounders, for example, race. Based on our experience, black people are highly segregated in the Pittsburgh region and clustered around the downtown area, one of the hot spots for air pollutants (Figure 27). Such a spatial coincidence between air pollutants and race groups may confound our estimation of pollutants' mortality risks.

5.5.3 Potential Exposure Misclassification

Due to the irregular shapes of the ZCTAs and the weakness of spatiotemporal Kriging, the exposure to air pollutants may be misclassified in our study domain. According to the spatial distribution of the air monitors, the probability of exposure misclassification is larger in the counties of Westmoreland and Washington compared to Allegheny County and the city of Pittsburgh area. To compare the health effects estimated from the data restricted to within the city of Pittsburgh, Allegheny County and all the three counties, we can briefly evaluate the potential exposure misclassification. If the health risks estimated from the three areas are consistent with each other, it suggests that the accuracies of exposure assessment are comparable between the three areas. Therefore, the exposure misclassification of the other two counties is at a similar level to Allegheny County (baseline). Otherwise, exposure misclassification is higher than that of Allegheny County. According to the regression results as shown in Figure 30, we suggest exposure misclassification may be relative higher than the baseline in the whole study domain for gaseous pollutants but equivalent to the baseline for the particulate matter.

5.5.4 Potential Confounding Effects

The six air pollutants are highly correlated with each other especially in the temporal dimension, which is shown by the empirical cross-variograms between each pair of pollutants as shown in Figure 32. To evaluate the confounding effects caused by the collinearity of air pollutants, we constructed two-pollutants models by adding the second pollutant to the regression models (known as baseline models) in Section 5.4.2. The two-pollutants models are displayed in Figure 33. For total mortality and mortality of cardiovascular diseases, in our spatiotemporal GEE models, the two-pollutants regression results are consistent with the baseline models, which suggest that after controlling spatiotemporal auto-correlations, air pollutants may not confound the health effects of each other. For mortality of cancer and respiratory diseases, the heath effects of $PM_{2.5}$, PM_{10} and O_3 are statistically significantly decreased after adding NO_2, which suggests that $PM_{2.5}$, PM_{10} and O_3 are potential confounders for NO_2. The results of the two-pollutants models suggests NO_2 to be a critical pollutant associated with mortalities of respiratory diseases and cancer in the Pittsburgh region, and indicates that traffic emission may be one of the most important sources of air pollutants in Pittsburgh, from 1999 to 2008.



If the two air pollutants are correlated or the same pollutants are auto-correlated in temporal dimension, the empirical variograms will increase with time lag, otherwise, the empirical variograms will be flat; if the two air pollutants are correlated or the same pollutants are auto-correlated in spatial dimension, the entire curve of variograms will point-wisely increase with the increase of spatial lag intervals, e.g. $\underline{\text{PM}_{10} \text{ vs } \text{O}_3}$, otherwise variogram curves will be similar between different spatial lag intervals.

Figure 32: Empirical variograms and cross-variograms of six air pollutants by temporal lag grouped by spatial lag.

We used the regression results for Allegheny County data to generate this figure. In the panel of plots, each row shares the same baseline pollutant and each column shares the same health outcome. For the x-axis of each plot, the first model is always a one-pollutant model, the same as those shown in Section 5.4.2.

Figure 33: Two-pollutants modeling results: increase of relative risk for mortalities per IQR of baseline pollutants and their 95% confidence intervals for different combinations.

5.6 CONCLUSION

In this section, we developed a parameter-driven spatiotemporal Poisson regression model for environmental epidemiology and applied a novel generalized estimating equations approach to estimate the regression coefficients. Even though our methods are limited in a few respects, they can avoid overestimating health effects when ignoring spatiotemporal auto-correlation. We illustrated our methods using a study to associate air pollutants to ZIP code level daily counts of mortalities in the Pittsburgh region from 1999 to 2008. The study indicated that NO_2 is a key pollutant in Pittsburgh and significantly associated with total mortality, respiratory diseases and cancer with relative risks per IQR of 1.0098 (1.0021-1.0176), 1.0387 (1.0136-1.0644), 1.0251 (1.0095-1.0411), respectively.

6.0 SPATIOTEMPORAL ASSOCIATING LUNG CANCER INCIDENCE TO $PM_{2.5}$ AND SMOKING IN THE STATE OF PENNSYLVANIA, 2001-2007

In this chapter, we propose to study the chronic health effects of Ozone and associate it with incidences of lung cancer and its subtypes identified from cancer registries. However, considering latency period of 20 years for lung cancer, we will have to interpolate to accommodate the missing values in the histological records for O_3 in 1980's using a nonparametric space-time optimization method. To adjust for the confounding effects of smoking, we will first decompose it into its spatial and temporal dimensions using another optimization method and then control its spatial and temporal patterns separately in our regression models. In the final step of statistical analysis, we will apply a typical Bayesian spatiotemporal hierarchical model to associate cancer incidences with exposure to O_3, smoking and other socioeconomic risk factors. This chapter aims to provide a comprehensive example to illustrate widely-used Bayesian hierarchical space-time models, and show how to construct a flexible convex optimization to solve problems in spatiotemporal analysis for epidemiological studies and in particular environmental epidemiological studies.

6.1 INTRODUCTION AND DATA

6.1.1 Introduction

Lung cancer is a key health issue in the Pittsburgh region. According to cancer statistics reports by the CDC [WONDER, 2014], the age adjusted incidence rate for lung and bronchial cancer in Pittsburgh are significantly higher than those in

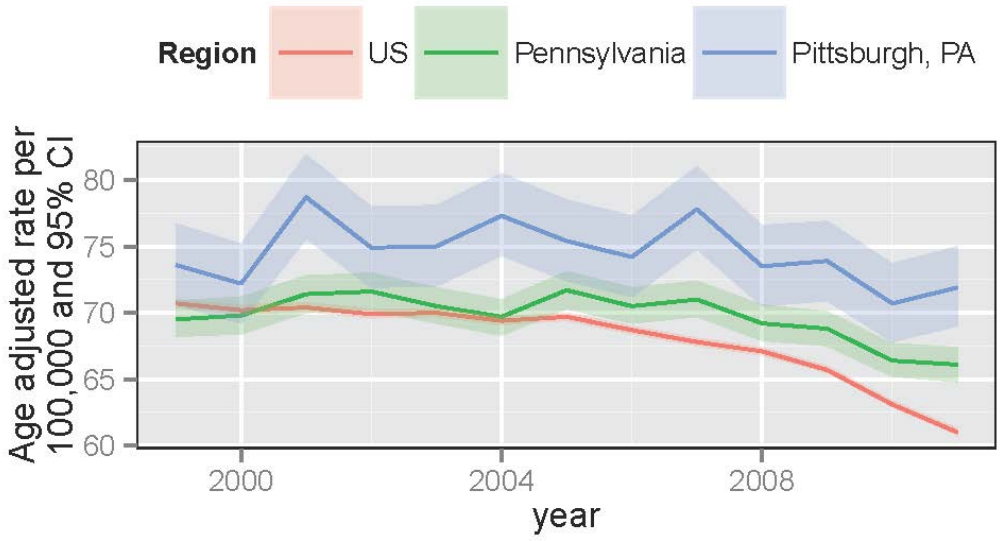


Figure 34: Age-adjusted incidence rates for lung and bronchial cancer per 100,000 population and 95% CIs for Pittsburgh, the state of Pennsylvania and the United states from 1999 to 2011.

the state of Pennsylvania and the United states (Figure 34). Lung cancer has been associated with a series of risk factors including smoking [Peto et al., 2000] and second hand smoking [Öberg et al., 2011], radon radiation [Darby et al., 2005], air pollutants [Beeson et al., 1998, Nyberg et al., 2000, Jerrett et al., 2009], other hazardous substances such as arsenic [Smith et al., 1998, Chen et al.,] and asbestos [Doll, 1993] and socioeconomic status [Mao et al., 2001]. Among all the risk factors, we hypothesize that air pollution contributes to the extremely high risk of lung caner in Pittsburgh and thus are motivated to design a ecological study to associate air pollutants with lung cancer incidence in the state of Pennsylvania. In this study, we will also adjust for county level indicators of other risk factors including smoking and socioeconomic status.

Risk factors for various subtypes of lung cancer differ. Taking smoking as an example, according to the previous study, even though smoking raises risks for all subtypes of lung cancer, it is more associated with small cell carcinoma than adenocarcinoma [Kenfield et al., 2008]. Kenfield et. al., (2008) has reported that an increase of one additional year of smoking is related with a 6% increase in the risk of adenocarcinoma, compared to 7% for large cell, 10% for squamous cell and 12% for small cell carcinoma. Therefore, in our ecological spatiotemporal study, we will also explore the difference in the effects of air pollutants on the four major subtypes of lung cancer.

Both chronic and acute exposure to air pollutants have been associated with indicators of respiratory diseases such as mortality [Pope III et al., 1995], hospital admissions [Dominici et al., 2006], lung function biomarkers [Brunekreef et al., 1997, Lin et al., 2011] and lung cancer [Beeson et al., 1998, Nyberg et al., 2000, Jerrett et al., 2009]. However, compared to acute health effects, identifying risk factors for chronic effects such as lung cancer incidence is usually more challenging due to the possibility of latency periods for chronic diseases. For lung cancer, the latency period of exposure to various air pollutants and smoking has been reported to be in a range of 20-30 years [Weiss, 1997, Nyberg et al., 2000]. Therefore, spatiotemporal studies of chronic diseases require retrospective records of risk factors, which usually are collected by questionnaire or histological records. Cancer registries such as Surveillance, Epidemiology, and End Results (SEER) and North American Association of Central Cancer Registries (NAACCR) provide comprehensive information about

cancer incidences and cancer mortalities, but have been rarely associated with air pollutants by spatiotemporal studies due to lack of large-scale and comprehensive histological records for air pollutants and other risk factors such as smoking. For example, the Air Quality System (AQS), the most comprehensive monitoring network maintained by US EPA, provides gaseous pollutants from the early 1980s, PM_{10} from the late 1980s and $PM_{2.5}$ from the 1990s. Even assuming a latency period of 20 years for lung cancer incidences, AQS can only characterize lung cancer risks for gaseous pollutants from 2000s and for particulate matter from the 2010s. Thus our ecological study will only associate a lag of 20 year for O_3 exposure with cancer incidences from 2001 to 2007. In addition, at the beginning period of monitoring air quality, the AQS data have in poor quality and contain a lot of missing values. Therefore, interpolating the missing values of histological records of air pollutants is another problem in spatiotemporal modeling of chronic diseases. We also have a similar problem when collecting data from smoking surveys. Therefore, in the follow sections, we will first develop spatiotemporal optimization problems to model O_3 and smoking data and then a Bayesian hierarchical model to regression risk factors an county level counts of lung cancer and its four subtypes.

6.1.2 Data Description

6.1.2.1 Cancer Registries Analogously to the cancer data in Section 2.2.3, lung cancer registries from 2001 to 2007 were also obtained from the Pennsylvania Department of Health, but were geocoded by counties rather than census tracts to avoid a large number of zero counts. The four major subtypes of lung cancer (adenocarcinoma, large cell, small cell and squamous cell carcinoma) were identified using the primary site code and the histology code for each records. For lung cancer and its subtypes, individual records were aggregated into yearly counts for each county of Pennsylvania according to diagnosis time and home address. Table 12 shows a brief summary of lung cancer counts by sex and year for the state of Pennsylvania. Finally our spatiotemporal study included 71,568 new cases of lung cancer and 31.5% of them were identified as adenocarcinoma, compared to 2.6% as large cell, 13.9% as small cell, 19.0% as squamous cell carcinoma and 32.9% as another subtype.

Table 12: Counts of lung cancer and its subtypes by sex and year groups for the state of Pennsylvania from 2001 to 2007.

Site/Sex	Counts							
	2001	2002	2003	2004	2005	2006	2007	Total
Bronchus & Lung								
Total	10022	10213	10079	10005	10330	10375	10544	71568
Male	5539	5603	5547	5436	5580	5609	5607	38921
Female	4483	4609	4532	4569	4750	4766	4937	32646
Adenocarcinoma								
Total	3106	3188	3159	3130	3265	3317	3382	22547
Male	1568	1588	1635	1601	1636	1650	1636	11314
Female	1538	1600	1524	1529	1629	1667	1746	11233
Large cell carcinoma								
Total	344	280	283	243	255	228	260	1893
Male	203	157	174	145	137	121	147	1084
Female	141	123	109	98	118	107	113	809
Small cell carcinoma								
Total	1427	1447	1396	1378	1430	1399	1452	9929
Male	743	807	707	689	732	698	711	5087
Female	684	640	689	689	698	701	741	4842
Squamous cell carcinoma								
Total	2036	1864	1915	1933	1927	1981	1969	13625
Male	1343	1219	1244	1250	1225	1296	1237	8814
Female	693	645	671	683	702	685	732	4811
Others								
Total	3109	3434	3326	3321	3453	3450	3481	23574
Male	1682	1832	1787	1751	1850	1844	1876	12622
Female	1427	1601	1539	1570	1603	1606	1605	10951

6.1.2.2 Smoking Data Behavioral Risk Factor Surveillance System (BREFSS) provides a comprehensive survey of adult smoking in the United States. However, the smallest spatial unit available in BREFSS is metropolitan statistical area (MSA), which usually consists of several counties. Therefore, we obtained estimated county level percent smoking data, 1996-2012 from a previous study [Dwyer Lindgren et al., 2014] and the long-term trend for adult smoking in the US, 1965-2011 from the CDC Website (http://www.cdc.gov/tobacco/data_statistics/tables/trends/cig_smoking/). Through comparing the trend data for the entire US and estimated results for Pennsylvania, we will decompose the county level percent smoking data into separate spatial and temporal trends using the solution to a convex optimization problem in the following sections. The time series of the long-term trend and county level estimators are displayed in Figure 35.

6.1.2.3 Demographic and Socioeconomic Data The county level demographic data by age, sex and race, and selected socioeconomic factors including median household income, percent of residences with education less than high school and commuting time were obtained from the 2000 census. The demographic data were used to adjust for the potential confounding effects of age, sex and race and calculate standardized incidence ratios (SIRs) analogously as in Section 2.2.3 in order to compare cancer risks between different counties. The SIR maps for lung cancer and its subtypes are shown in Figure 36 and the socioeconomic maps are shown in Figure 37. Among the selected socioeconomic factors, we highlighted commuting time, which may be an indicator for exposure to traffic emissions and thus potentially confound the effects of air pollutants on health.

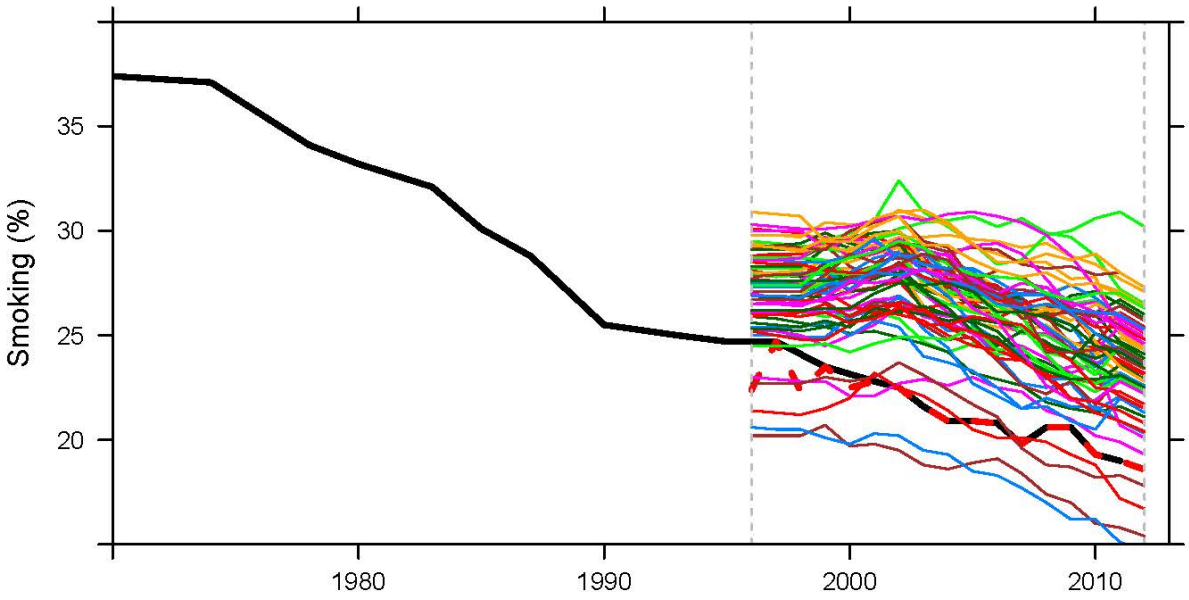
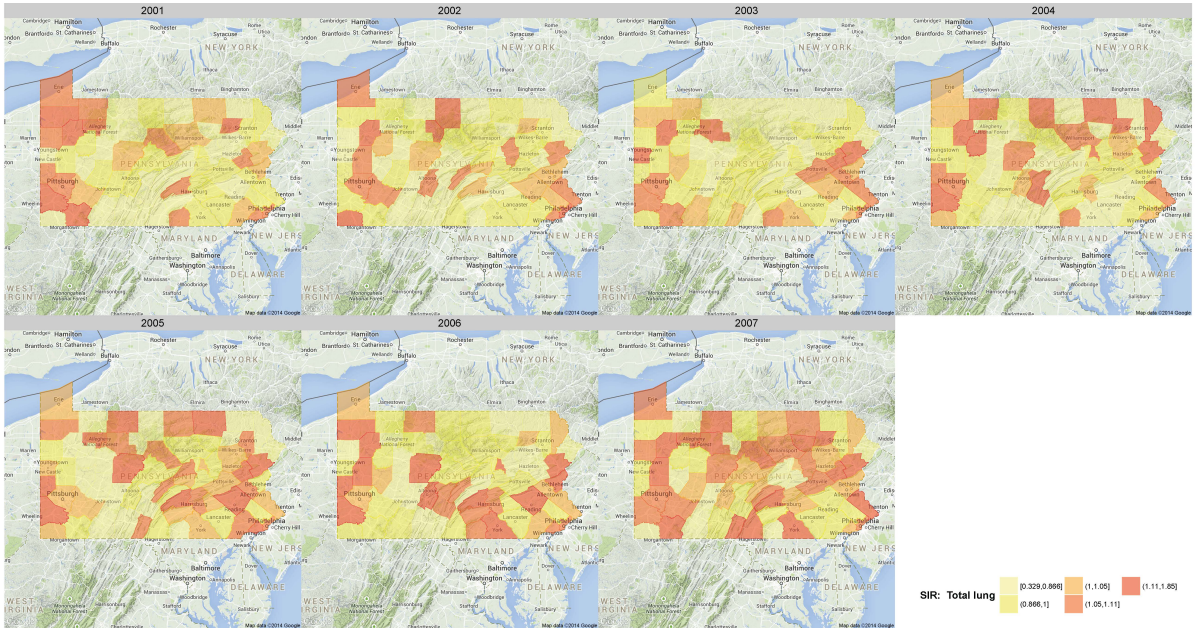
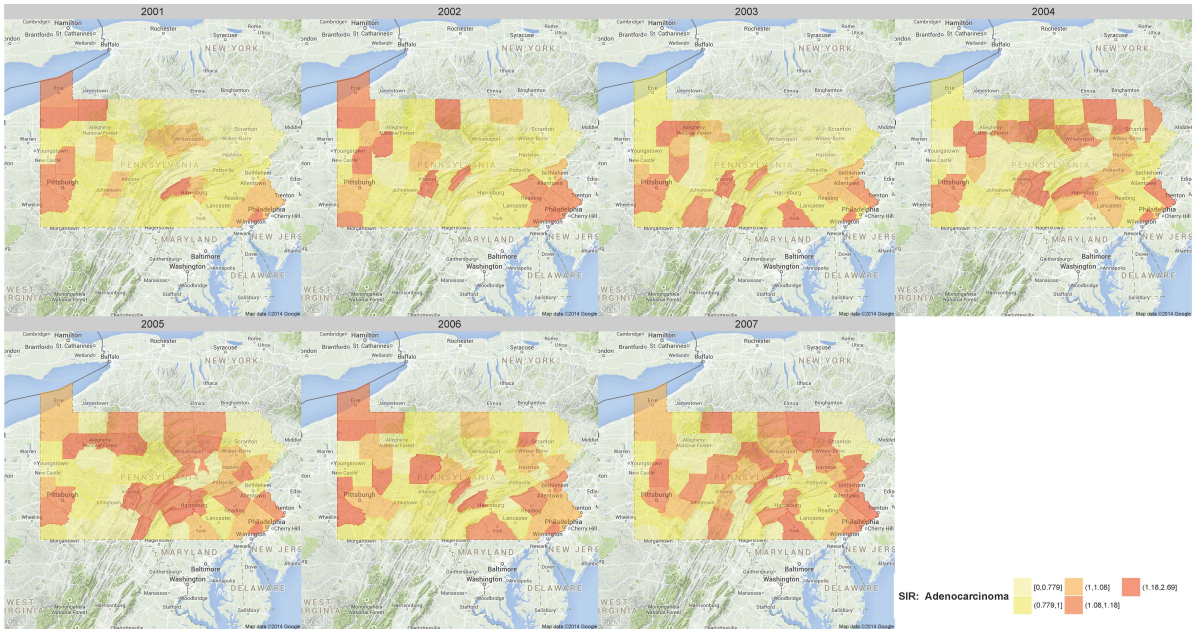


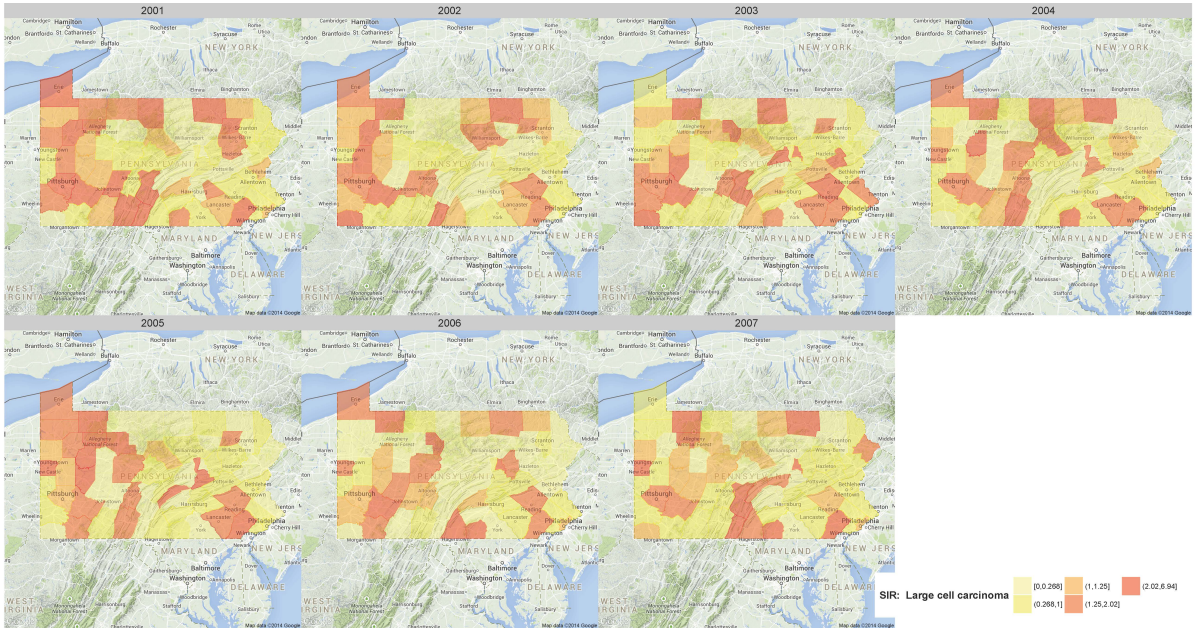
Figure 35: Long-term trends of current adult smoking in the United States, 1965–2012 (black solid line), time series of estimations of percent of smoking for all Pennsylvania counties (colored solid lines) and fitted long-term trends by spatiotemporal optimization (red dashed line).



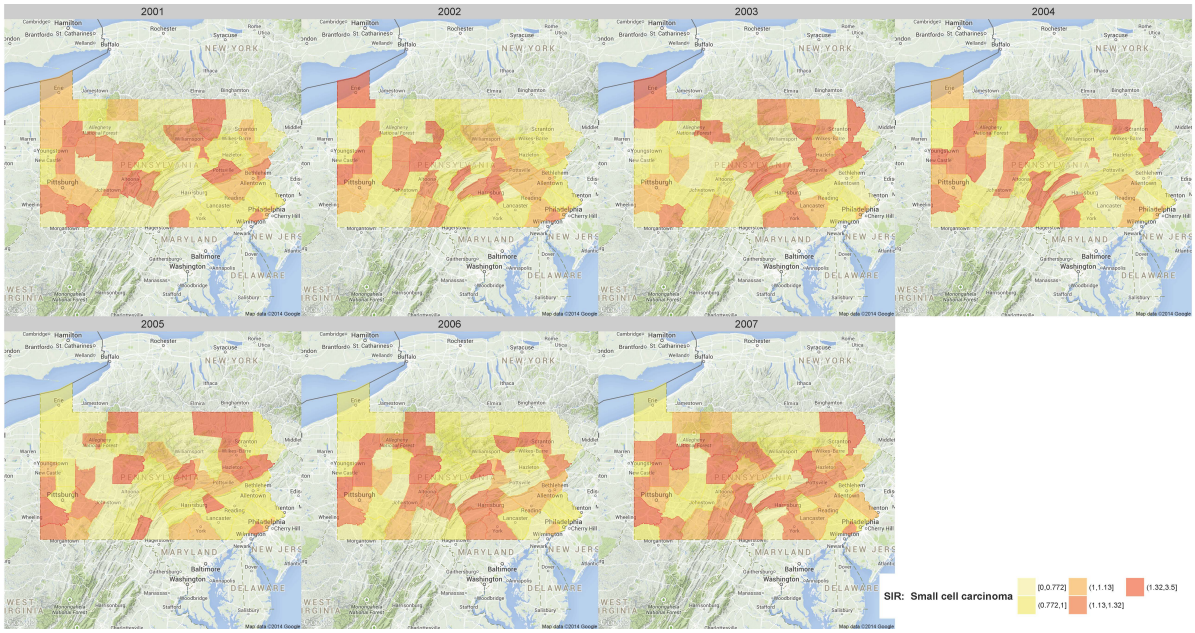
(a) Lung & bronchus



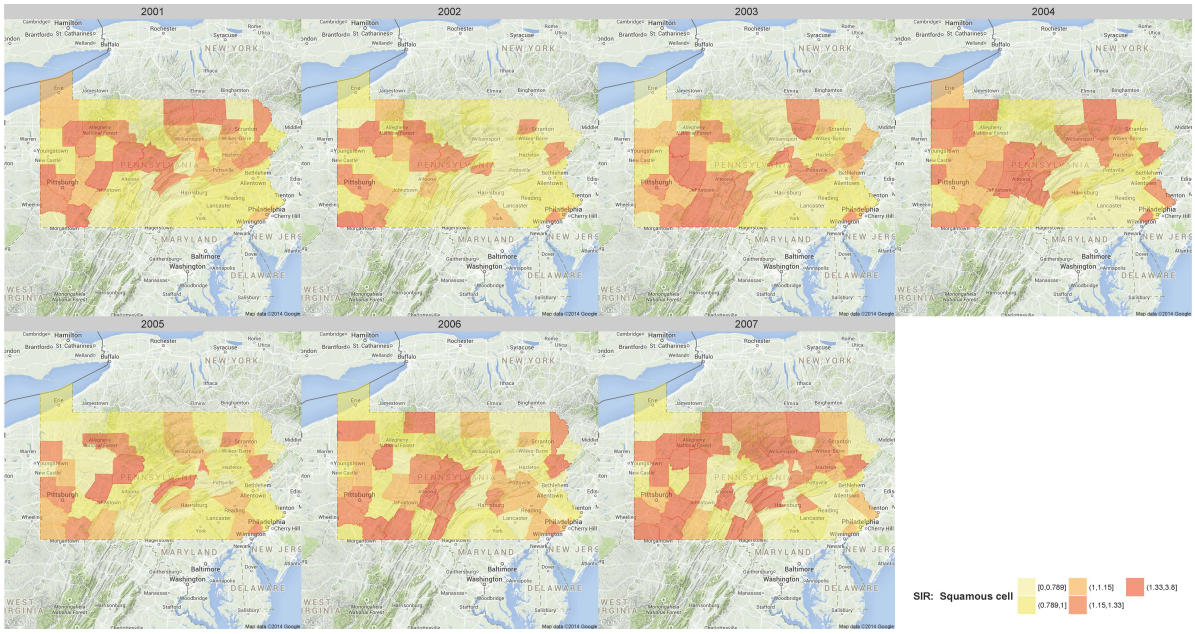
(b) Adenocarcinoma



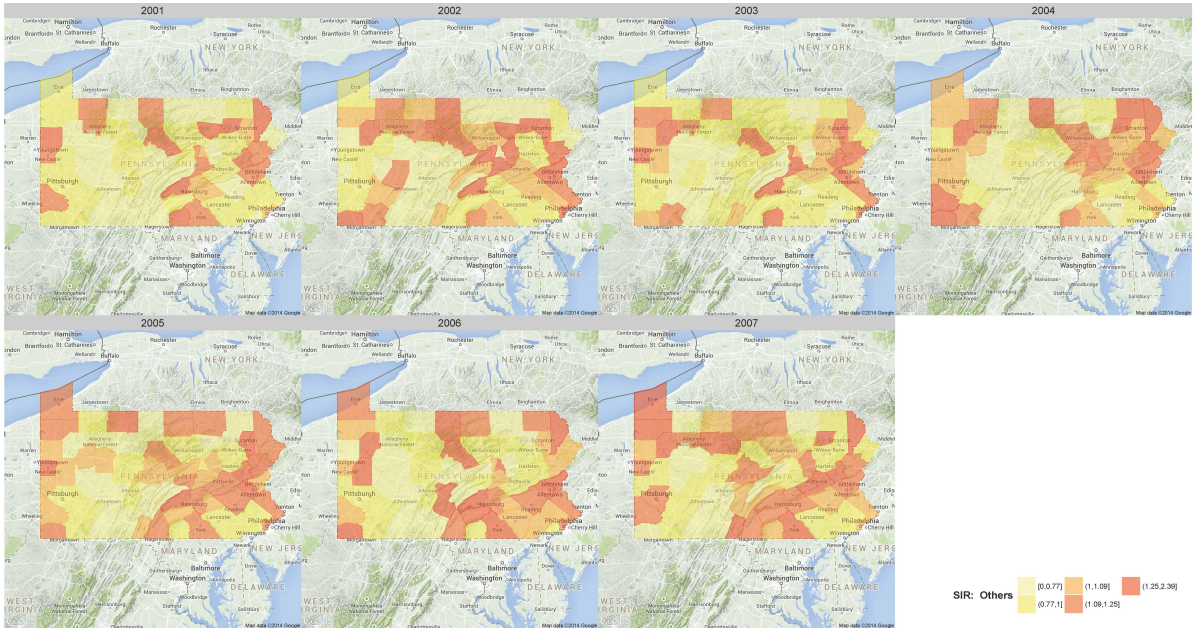
(c) Large cell carcinoma



(d) Small cell carcinoma

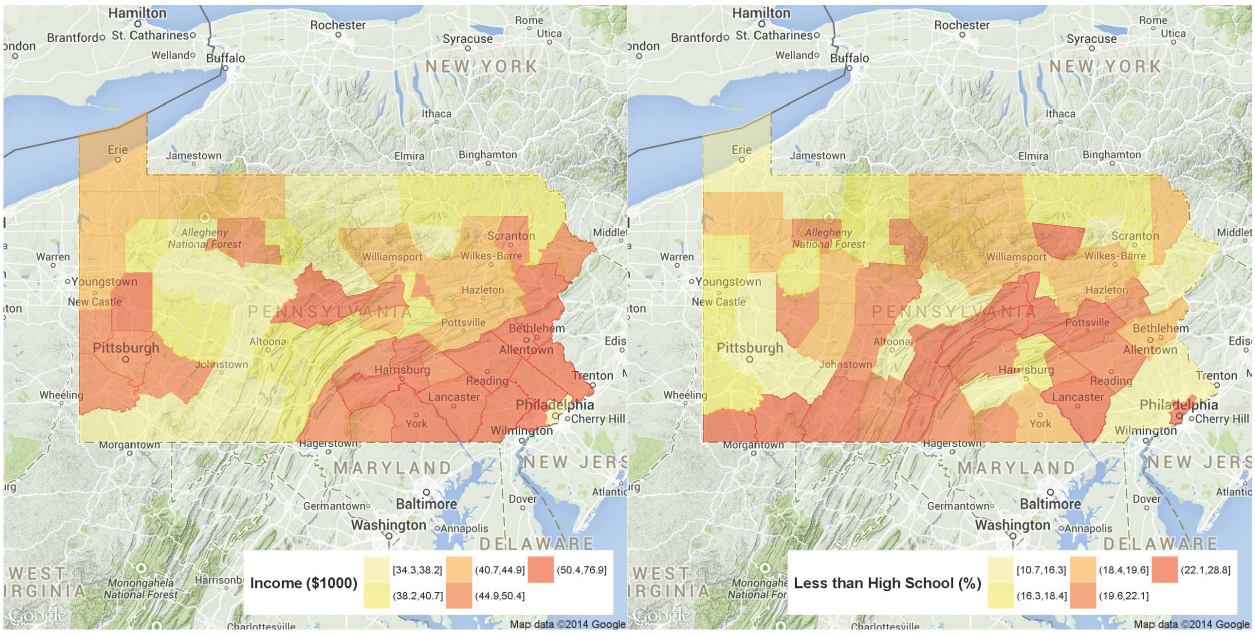


(e) Squamous cell carcinoma

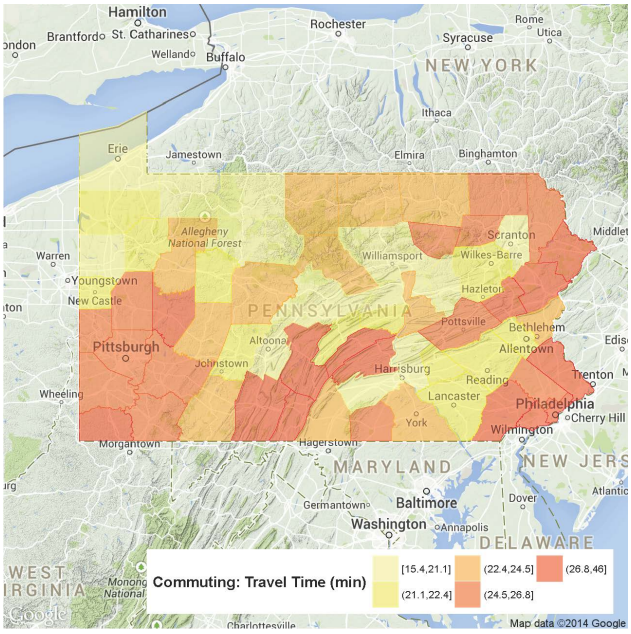


(f) Others

Figure 36: Standardized incidence ratios for lung cancer and its subtypes in all counties of Pennsylvania for 2001-2007.



(a) Income (b) Education



(c) Commuting

Figure 37: Selected county level socioeconomic factors: median household income (a), percent of residences with education less than high school (b), and commuting time (c) in the state of Pennsylvania for 2010.

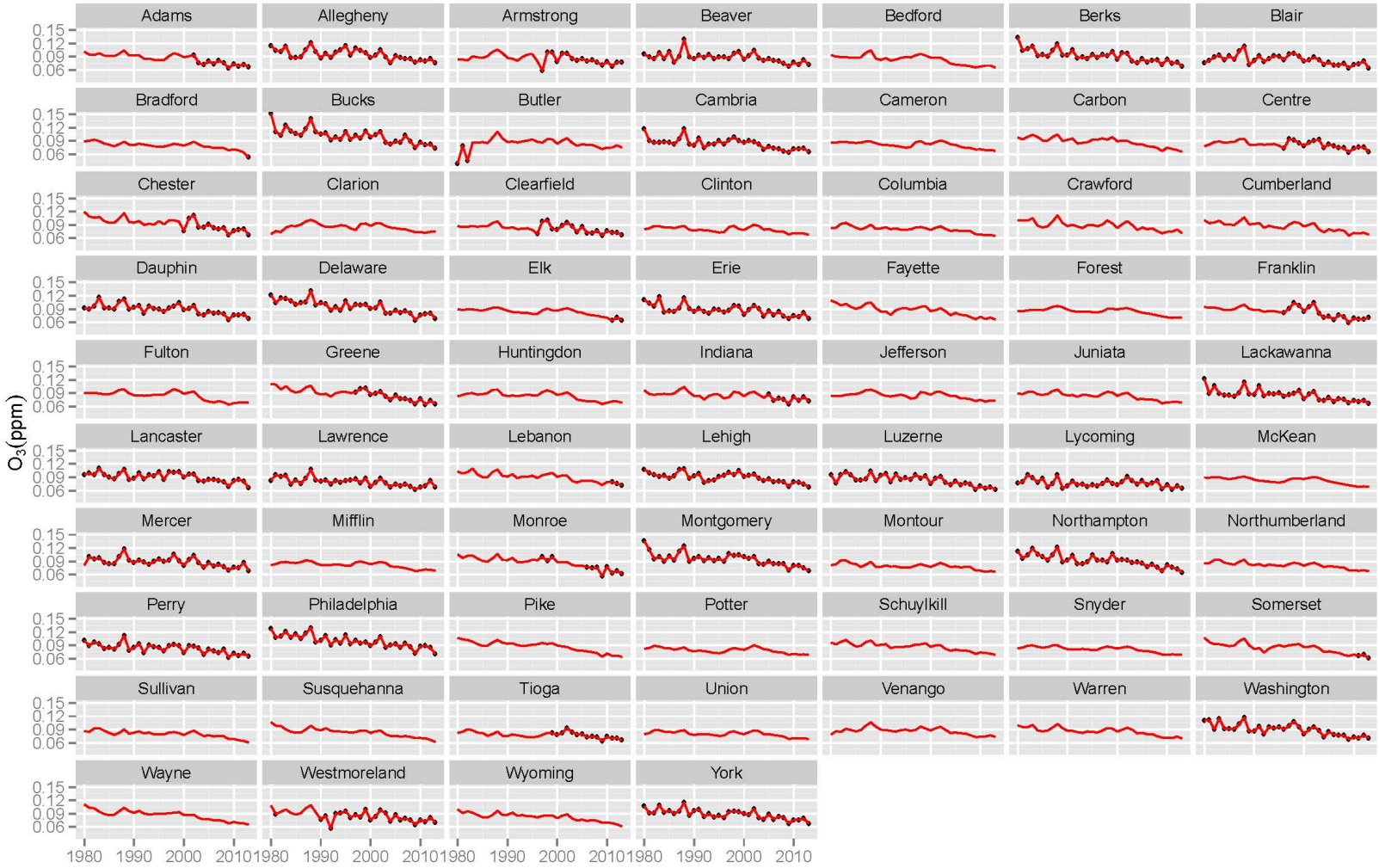


Figure 38: Annually 4^{th} maximum of 8-hour averages of O_3 (black dots) and interpolated time series by spatiotemporal optimization (red lines) for all 67 counties in the state of Pennsylvania for 1980-2013.

6.1.2.4 O_3 Data County level reports of yearly 4^{th} maximum of 8-hour averaged values of O_3 of air quality system (AQS) network were published on the US EPA website (http://www.epa.gov/airquality/airdata/ad_rep_con.html) from 1980. We obtained the annual reports from 1980 to 2013 and the time series of O_3 for all Pennsylvania counties are displayed in Figure 38. The figure shows that histological records for O_3 only cover 38 out of the 67 counties in Pennsylvania and for some counties, such as Chester and Clearfield, monitors were set up only from the late 1990s. Therefore, predicting large numbers of missing values is an important problem in our study.

6.2 STATISTICAL MODEL: SPATIOTEMPORAL OPTIMIZATIONS AND BAYESIAN HIERARCHICAL MODEL

6.2.1 Spatiotemporal Decomposition of Smoking Data

Assuming a latency period of twenty years for lung cancer, our spatiotemporal study requires county level smoking data from 1981 to 1987. In order to infer the unavailable histological data for smoking, we assume that the spatiotemporal pattern for county level smoking data in Pennsylvania can be decomposed into a constant spatial trend and a temporal trend, which is linearly associated with the long-term trend in entire US. Let $z_{s,t}$ denote the percent smoking at the t^{th} year and the s^{th} county of Pennsylvania and y_t denote percent smoking at the t^{th} year of the long-term trend in the entire US. Assuming two latent vectors $\boldsymbol{u} \equiv [u_1, \cdots, u_n]'$ and $\boldsymbol{v} \equiv [v_1, \cdots, v_m]'$ to denote the latent vectors for spatial and temporal patterns of smoking, respectively in Pennsylvania, we can construct a decomposition model as

$$z_{s,t} = u_s + v_t + \epsilon_{s,t}, \tag{6.1}$$

where $\epsilon_{s,t}$ is an error term. Assuming no scale bias between $z_{s,t}$ and y_t, the linear relationship between the county level and the long-term trend of smoking can be simplified as

$$y_t = \beta_0 + v_t + \xi_t. \tag{6.2}$$

Taking the above two Equations 6.1 and 6.2 and the spatial and temporal smoothness of the latent vectors ($\boldsymbol{u}$ and $\boldsymbol{v}$) into consideration simultaneously, we can construct an optimization problem as follows:

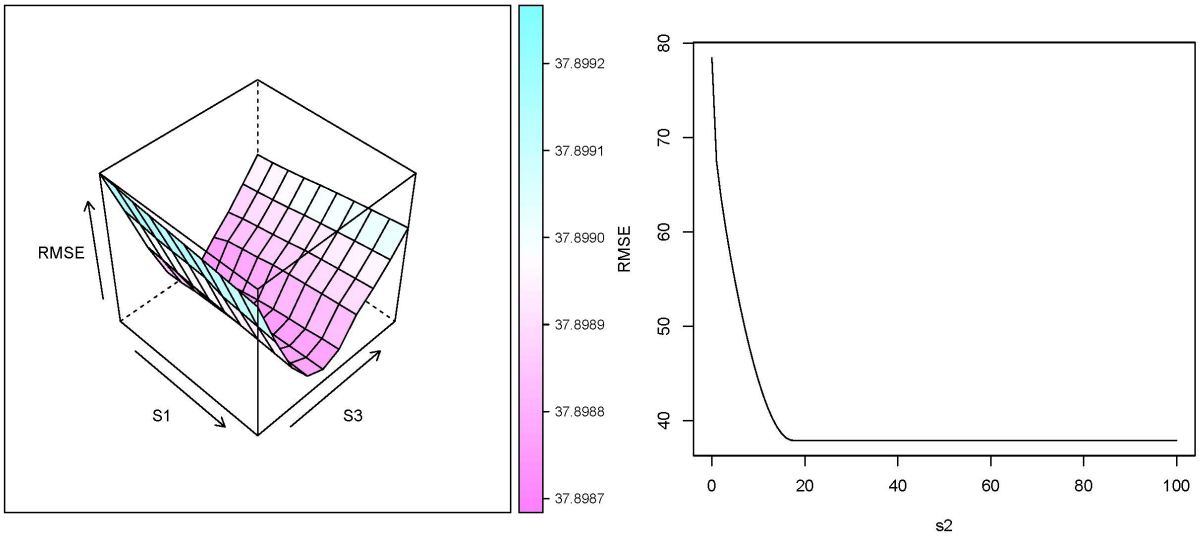
$$\begin{aligned} \min \quad & \|\boldsymbol{z} - \boldsymbol{I_u}\boldsymbol{u} - \boldsymbol{I_v}\boldsymbol{v}\|_2 \\ \text{subject to} \quad & \|\boldsymbol{y} - \boldsymbol{v} - \beta_0\|_2 < s_1 \\ & \|(\boldsymbol{I} - \boldsymbol{W_u})\boldsymbol{u}\|_2 < s_2 \\ & \|(\boldsymbol{I} - \boldsymbol{W_v})\boldsymbol{v}\|_2 < s_3, \end{aligned} \tag{6.3}$$

where $\boldsymbol{I_u}$ and $\boldsymbol{I_v}$ are indicator matrices for the spatial and temporal coordinates of $\boldsymbol{z}$; $\boldsymbol{W_u}$ and $\boldsymbol{W_v}$ are neighboring weights matrices similar to that of the SAR model in Section 2.2.1; s_1, s_2 and s_3 are tuning parameters and can be selected by cross-validation. This optimization is equivalent to a hierarchical model:

$$\begin{gathered} z_{s,t}|u_s, v_t \sim N(u_s + v_t, \epsilon_{s,t}),\ s = 1, \cdots, n,\ t = 1, \cdots, m \\ \Big[u_1, \cdots, u_n\Big]' \equiv \boldsymbol{u} \sim N_n\Big(0, (\boldsymbol{I} - \boldsymbol{W_u})^{-1}\boldsymbol{\Lambda}_u(\boldsymbol{I} - \boldsymbol{W_u})^{-T}\Big) \\ \Big[v_1, \cdots, v_m\Big]' \equiv \boldsymbol{v} \sim N_n\Big(0, (\boldsymbol{I} - \boldsymbol{W_v})^{-1}\boldsymbol{\Lambda}_v(\boldsymbol{I} - \boldsymbol{W_v})^{-T}\Big) \\ \boldsymbol{y}|\boldsymbol{v} \sim N(\boldsymbol{v} + \beta_0, \boldsymbol{\xi}),\ \text{where } \boldsymbol{\xi} \text{ is a diagonal matrix.} \end{gathered} \tag{6.4}$$

The inference for the hierarchical model can be made using MCMC or the likelihood method, in which estimating the variance-covariance parameters usually requires complicated computing efforts. By contrast in the spatiotemporal optimization, we select the tuning parameters using cross-validation and avoid this computational burden in the hierarchical model.

In the cross-validation, we left out each year's data iteratively, created a fine grid of the various values of the three tuning parameters and selected a set that minimized the cross-validation error (RMSE), which is shown in Figure 39. In the cross-validation analysis, we found that RMSE decreased to a constant level with an increase of s_2 at any fixed values of s_1 and s_3. The tuning parameters s_1, s_2 and s_3 were selected as 0, 20 and 4.22. $s_1 = 0$ represents an extreme restriction on the first constraint, and explains why the fitted long-term trend ($\boldsymbol{v} + \beta_0$) is exactly equal to the observed one in Figure 35. The fitted spatial pattern ($\boldsymbol{u}$) is mapped in Figure 40 and explains 68% of the variance of the observed county level data $\boldsymbol{z}$, compared to 16% for the temporal trend ($\boldsymbol{v}$) and 16% for the residuals ($\boldsymbol{\epsilon}$). Accordingly, this suggests that the spatial heterogeneity of smoking is an important issue in



(a) RMSE by s_1 and s_3

(b) RMSE by s_2

Figure 39: Cross-validation errors by different values of tuning parameters.

the state of Pennsylvania and may be more related with lung cancer risks than the temporal trend of smoking.

6.2.2 Spatiotemporal Interpolation of O_3 Data

Interpolating large numbers of missing values is a critical problem to be solved in exposure assessment of O_3. Analogously to the last section, we will take the spatiotemporal smoothness of O_3 and its measurement errors into consideration and develop an optimization problem as follows:

$$\begin{aligned} \min \quad & \|(\boldsymbol{I}-\boldsymbol{W_s})\boldsymbol{x}\|_2 \\ \text{subject to} \quad & \|\boldsymbol{y}-\boldsymbol{I_y}\boldsymbol{x}\|_2 < s_1 \\ & \|(\boldsymbol{I}-\boldsymbol{W_t})\boldsymbol{x}\|_2 < s_2 \\ & \|(\boldsymbol{I}-\boldsymbol{W_{st}})\boldsymbol{x}\|_2 < s_3, \end{aligned} \tag{6.5}$$

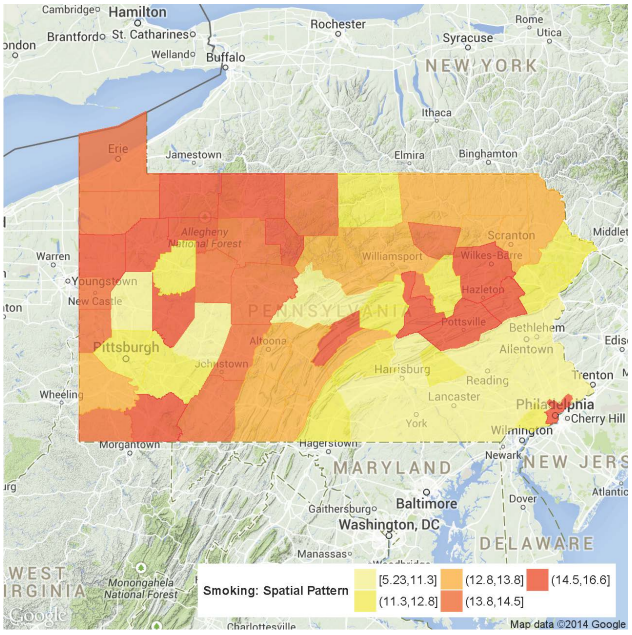


Figure 40: Fitted constant spatial pattern for adult smoking for all counties of Pennsylvania.

where $y_{s,t}$ and $x_{s,t}$ denote O_3 measurements and the latent true value at the t^{th} year and the s^{th} county of Pennsylvania with vector form $\boldsymbol{y}_m$, I_y denotes a $m \times n$ matrix to identify measurements and W_s, W_t and W_{st} denote spatial, temporal and cross-spatiotemporal (measurements at neighboring areas but at different time points) neighboring weights matrices. s_1, s_2 and s_3 are tuning parameters to control O_3 monitoring errors, temporal smoothness and cross-spatiotemporal smoothness, respectively. Similar to the last section, the spatiotemporal optimization of O_3 can also be converted into a hierarchical model.

In this optimization, we arrange measuring errors as a constraint instead of an objective function in order to control the interpolated values flexibly. Traditional interpolation methods, e.g. spline and LOESS generate fitted values at both missing and observed points of a time series or spatial field and thus provide confusing choices at observed points for further usage. (Should we use all the interpolated values or only use them at the missing points in further analysis?) In order to avoid this scenario, we manually restrict s_1 to be zero to make sure that the interpolated values at the observed points are exactly equal to the observed

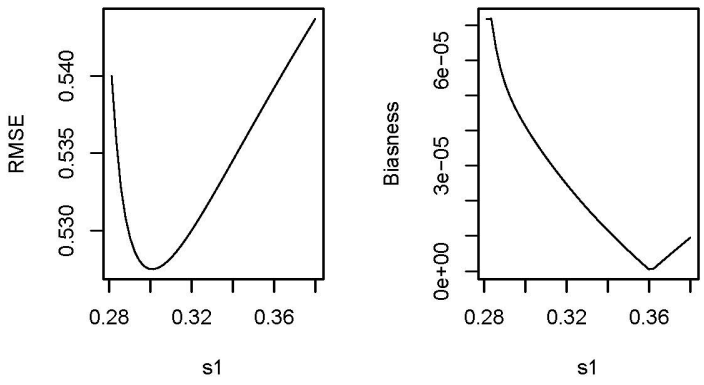


Figure 41: 10-fold cross-validation error (left) and biasness (right) for tuning parameter s_2.

values as shown in Figure 38. In addition, in practice to simplify the cross-validation analysis, we ignored the cross-spatiotemporal smoothness by setting s_3 to be infinity. Therefore, the only tuning parameter left is s_2, which was selected by minimizing the 10-fold cross-validation error (RMSE) as shown in Figure 41. The interpolated values are mapped in Figure 42.

6.2.3 Spatiotemporal Bayesian Hierarchical Model

Applying the aforementioned methodology in Section 2.2.2, we constructed spatiotemporal hierarchical models to regress yearly and county level counts of lung cancer and its subtypes on risk factors including O_3, percent smoking, income, education and commuting time and an offset of sex, age, race adjusted expected counts calculated using the demographic data. However, considering seven temporal points in the spatiotemporal model, we use a fixed temporal effect with 6 degrees of freedom instead of an AR(1) random effect to reduce the computational effort for the hierarchical models. The models were fitted using the R package `CARBayes`. We also calculated Deviance information criterion (DIC) to compare those models.

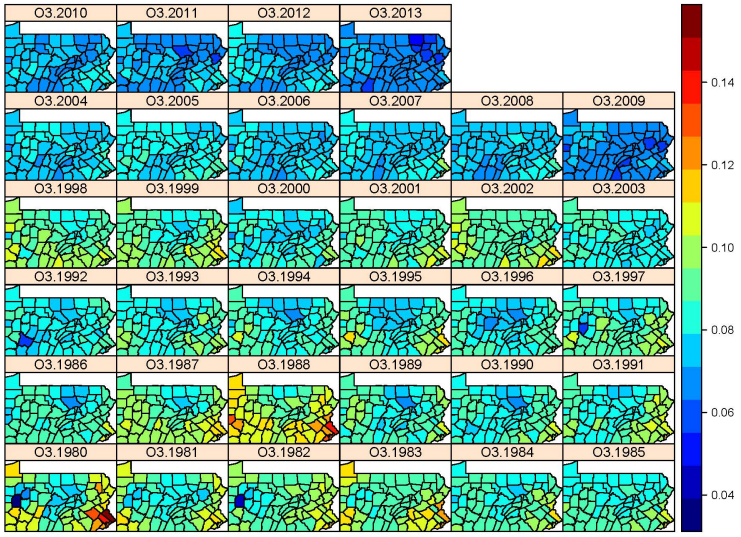


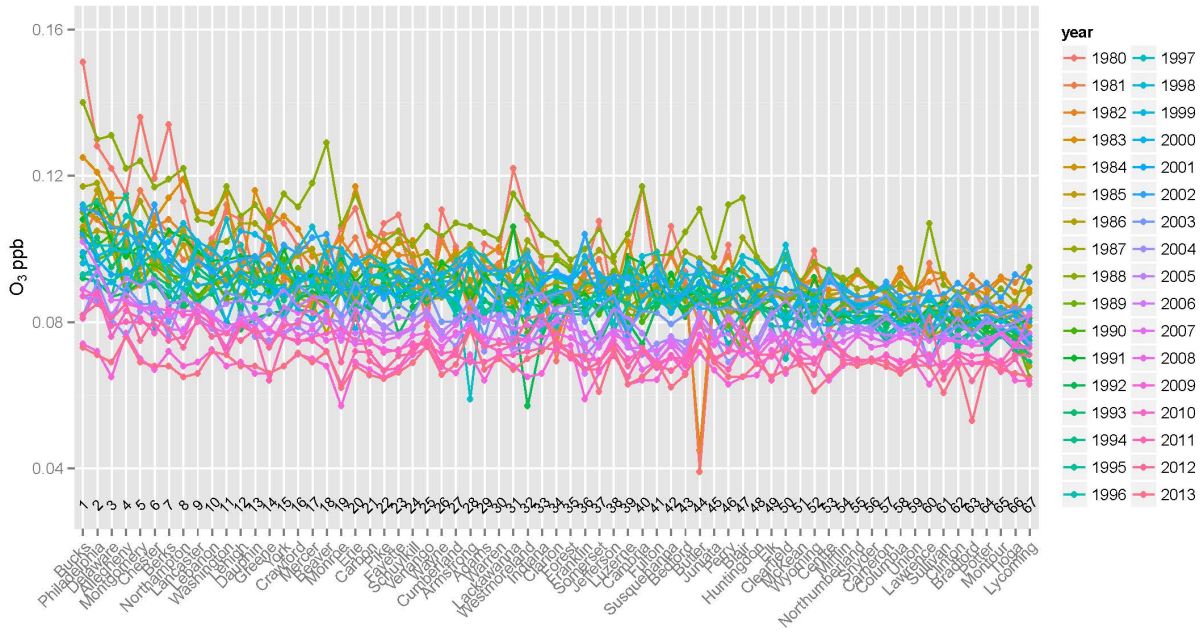
Figure 42: Interpolated O_3 maps for all counties in Pennsylvania for 1980-2013.

6.3 RESULTS

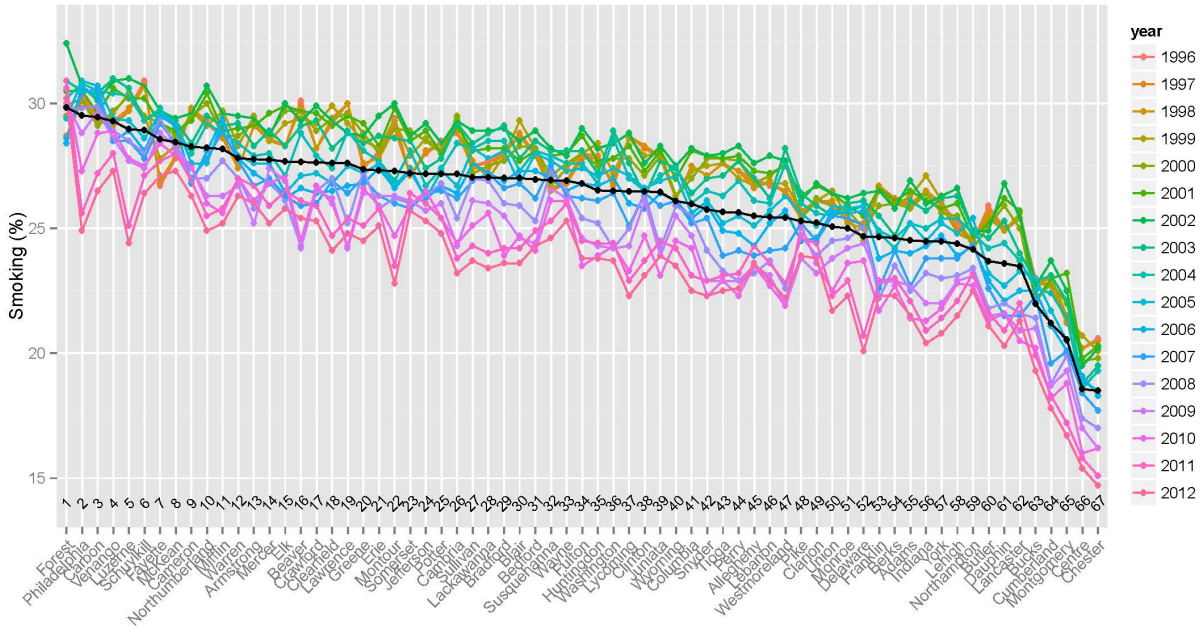
6.3.1 Descriptive Analysis

In order to compare O_3 exposure, smoking and risks of cancers between different counties in the state of Pennsylvania, we plot the values by their spatial patterns in figures 43 and 44. Among all the counties, SIR of Allegheny is ranked as the 8^{th} highest for total lung cancer, compared to the 9^{th} for adenocarcinoma, the 16^{th} for small cell, the 10^{th} for large cell, the 16^{th} for squamous cell carcinoma and the 36^{th} for other subtypes. While Allegheny County is ranked as the 4^{th} highest for O_3 exposure but as the 45^{th} highest for smoking. The descriptive analysis of the spatial trends suggests that air pollutants and not the smoking are the potential causes for the high risk of lung cancer in Allegheny County.

6.3.2 Regression Results

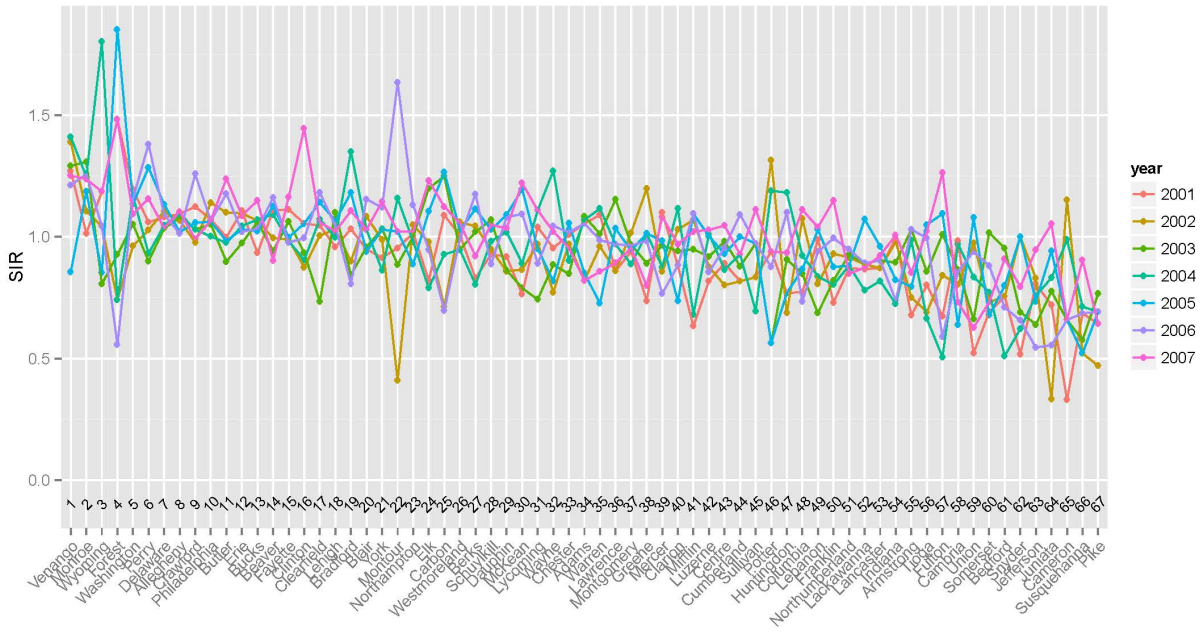


(a) Interpolated O_3 for 1980-2013

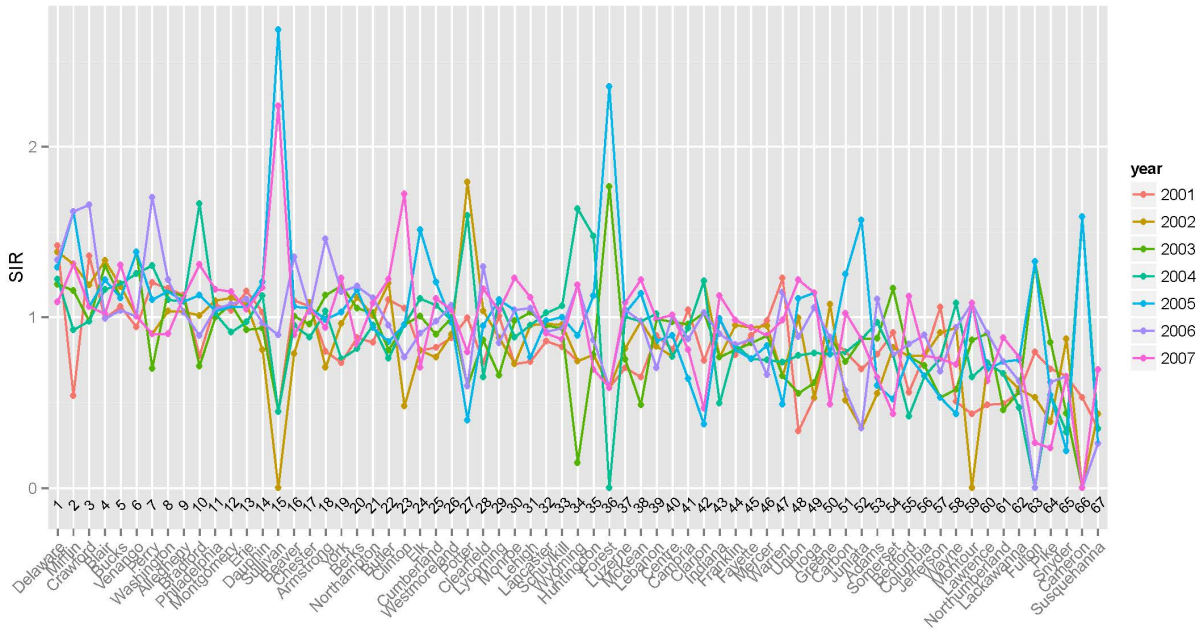


(b) Estimated smoking for 1996-2012 [Dwyer Lindgren et al., 2014]

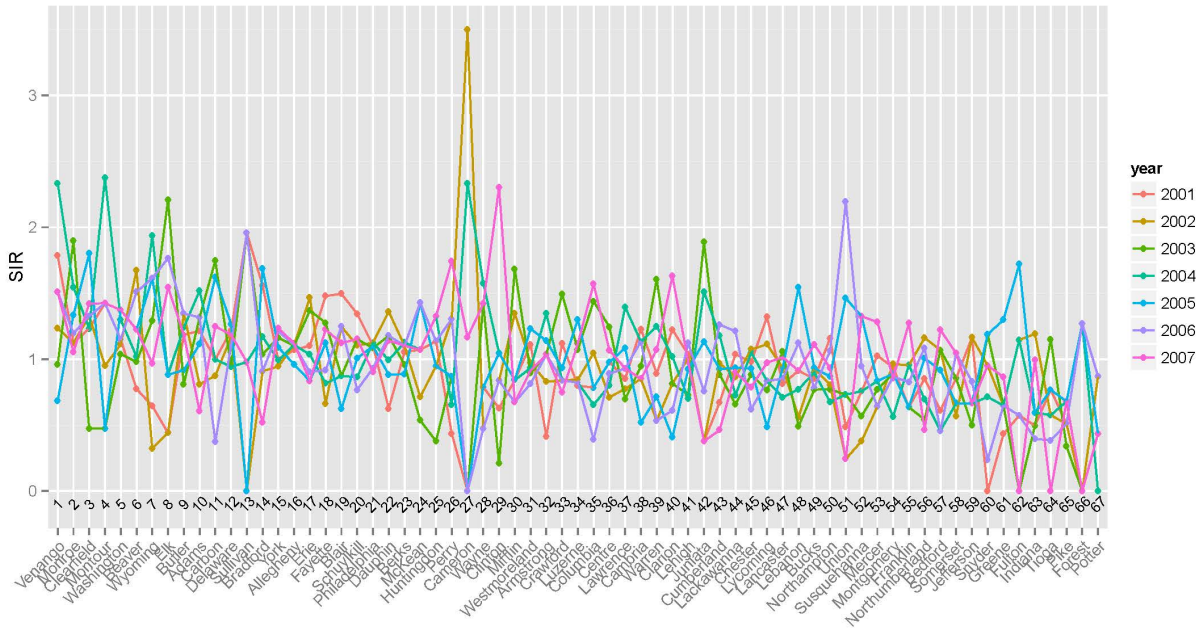
Figure 43: Plots of O_3 for 1980-2013 (a) and estimated smoking for 1996-2012 and (b) by the ranks of the averaged spatial patterns.



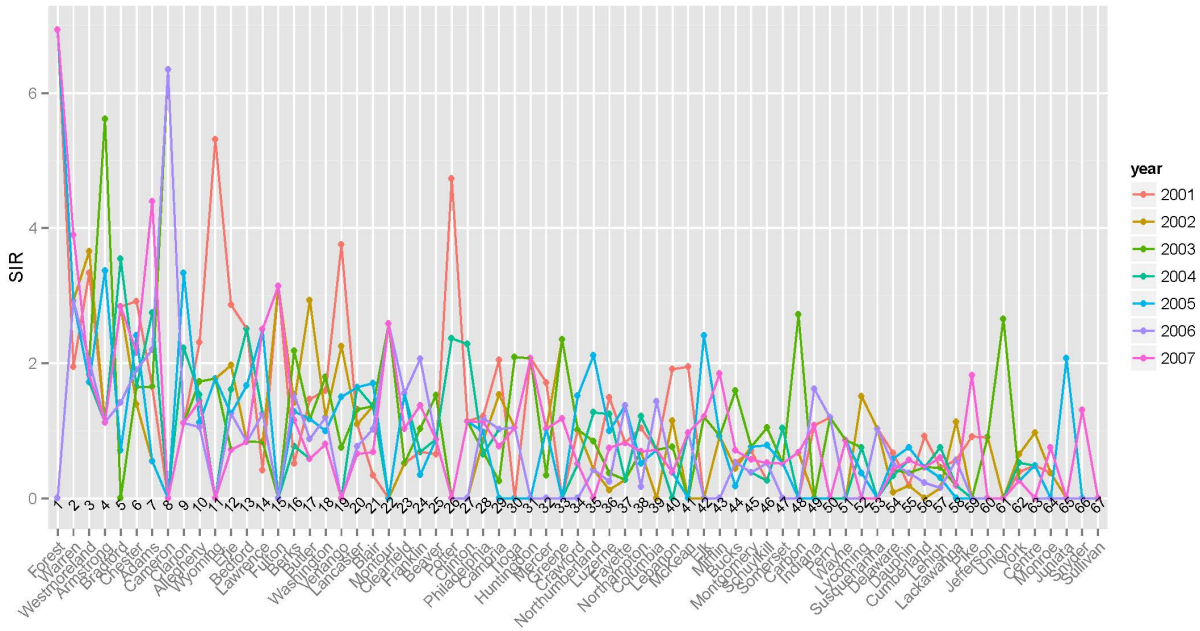
(a) Total lung cancer



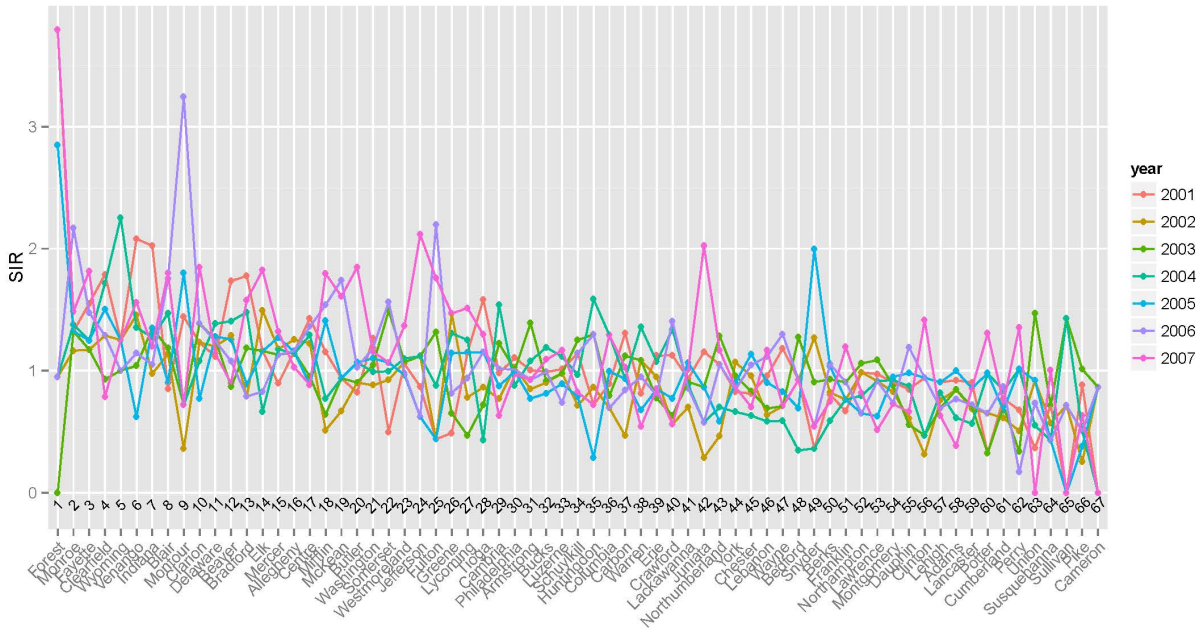
(b) Adenocarcinoma



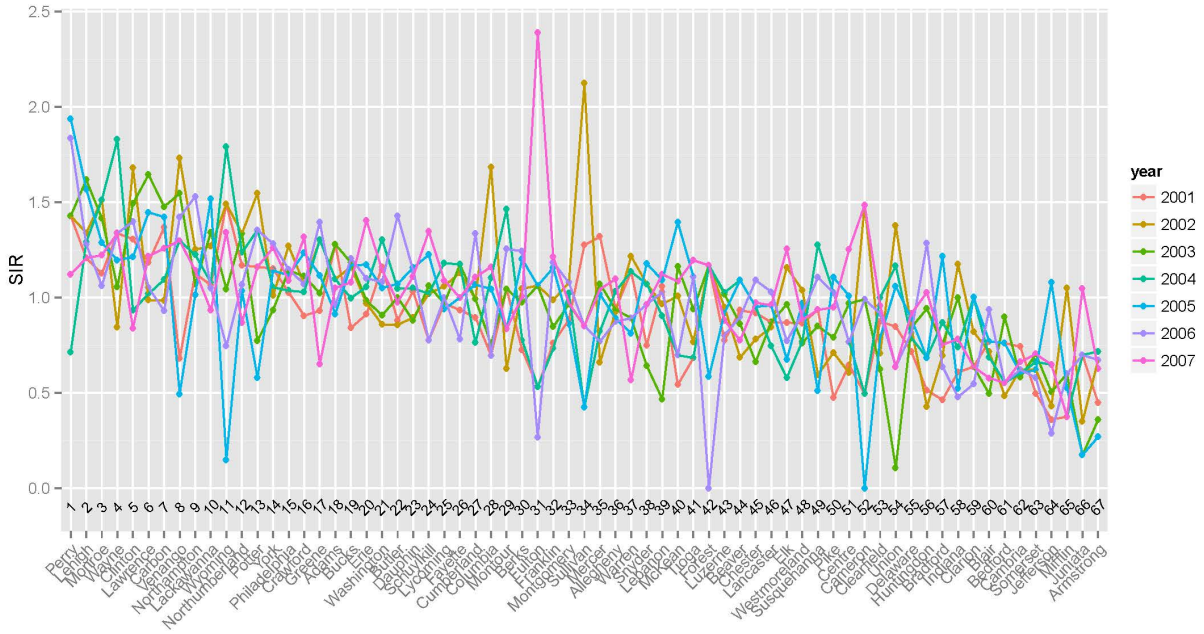
(c) Small cell carcinoma



(d) Large cell carcinoma

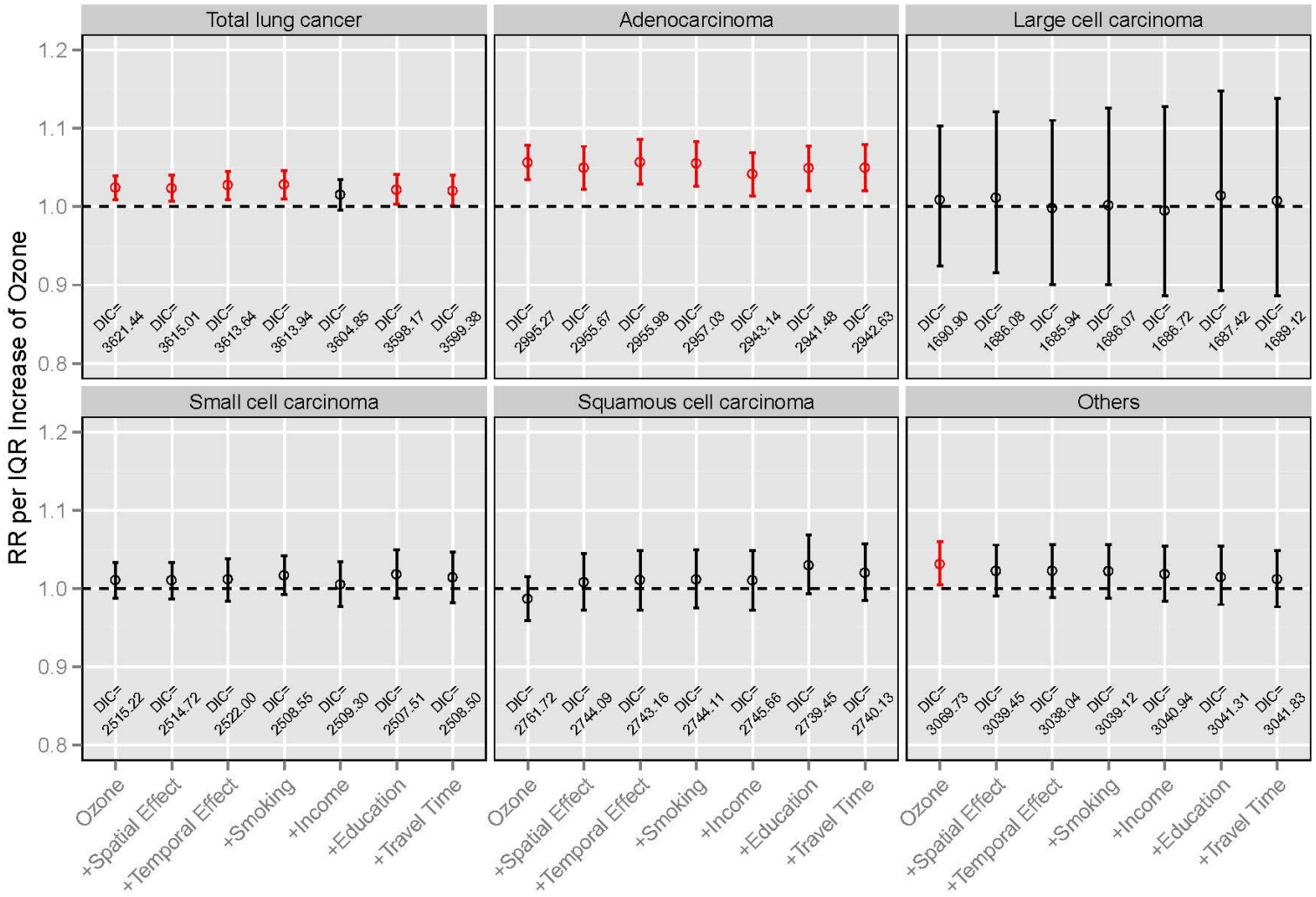


(e) Squamous cell carcinoma



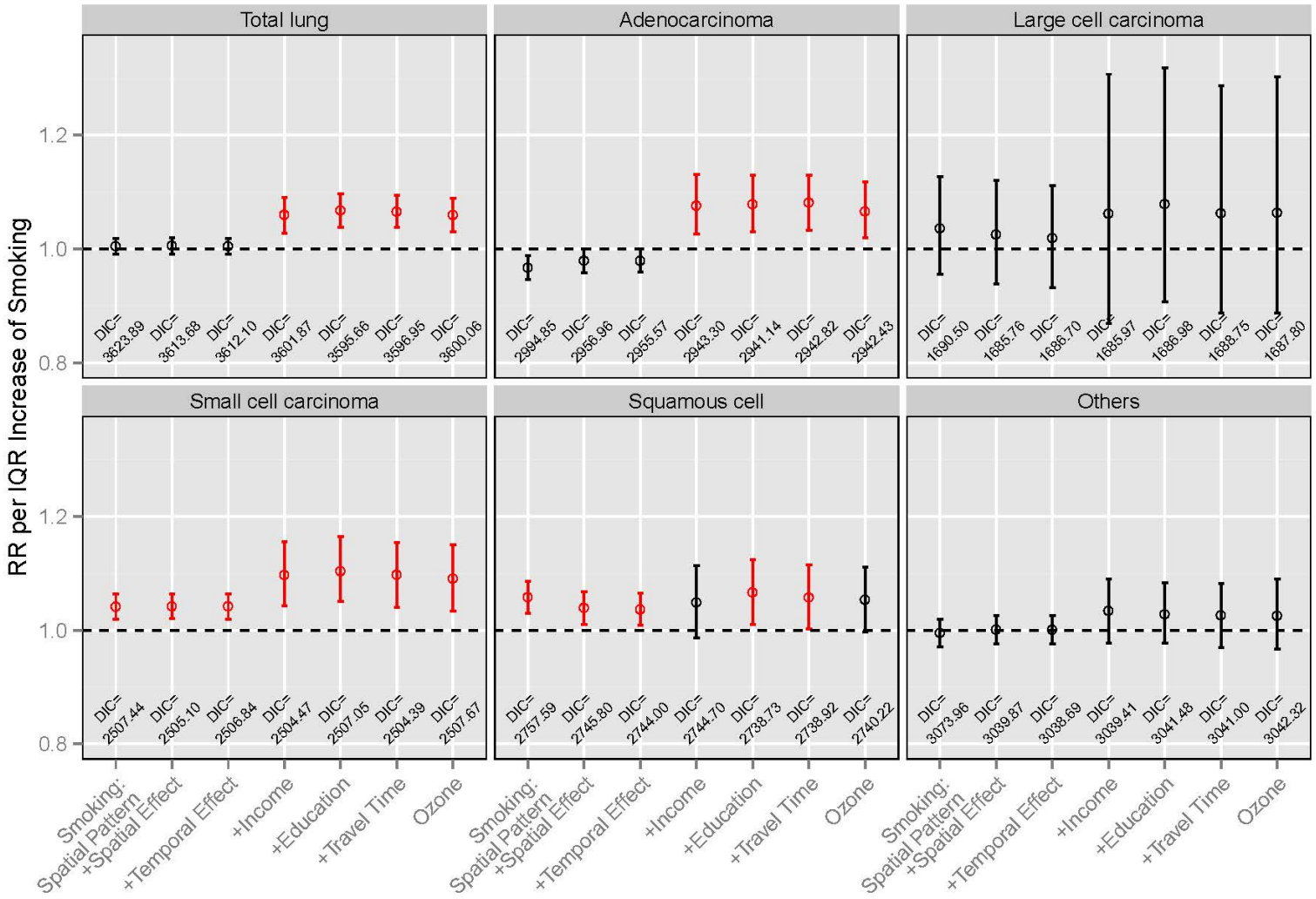
(f) Others

Figure 44: Plots for SIRs of lung cancer (a) and its subtypes (b,c,d,e,f) by the ranks of averaged spatial patterns.



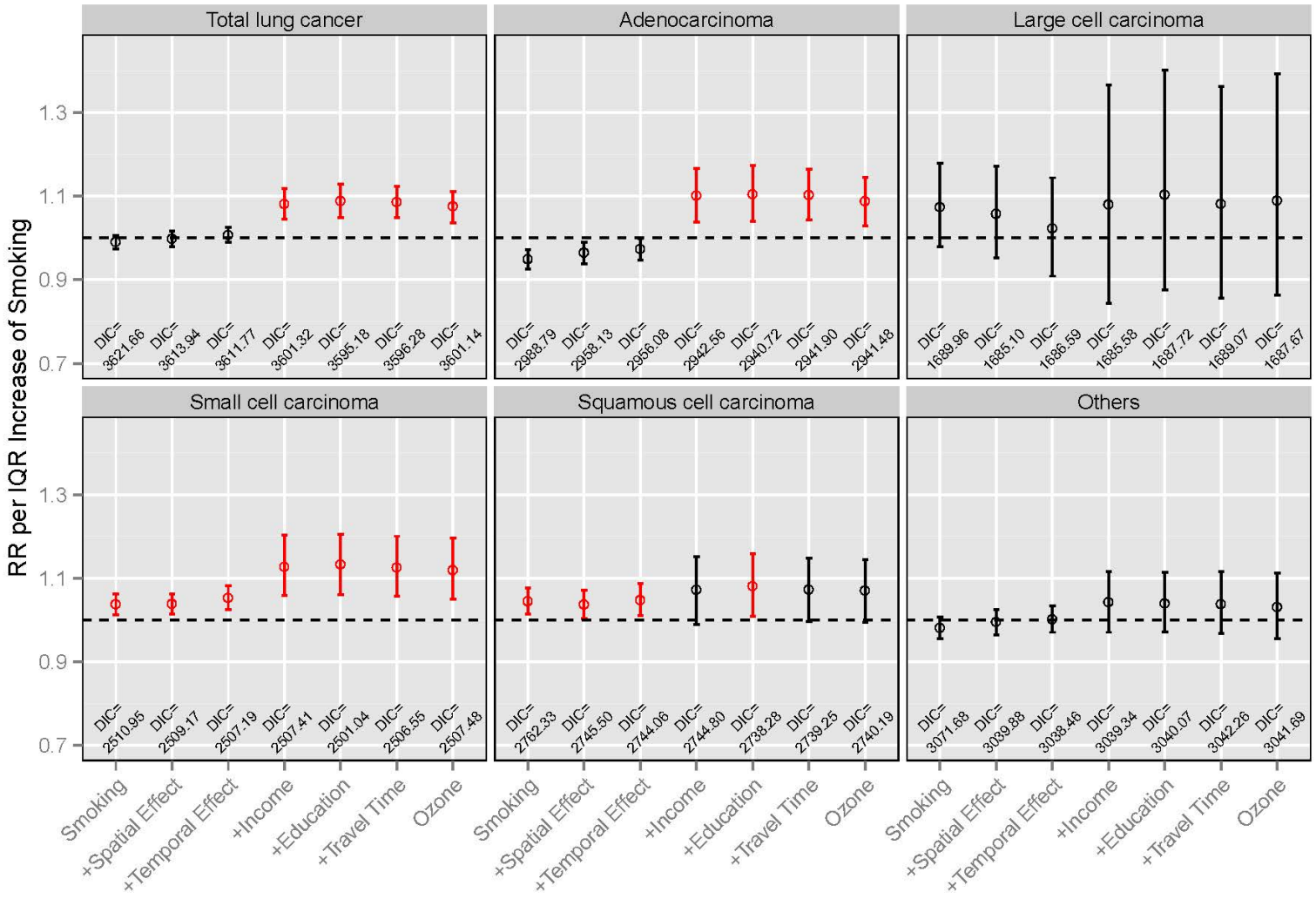
The significantly positive relative risks are highlighted by red color and the DICs are listed for each model.

Figure 45: Relative risks and their 95% CIs per IQR increase in O_3 exposure for lung cancer and its subtypes using sequentially adjusted models.



The significantly positive relative risks are highlighted by red color and the DICs are listed for each model.

Figure 46: Relative risks and their 95% CIs per IQR increase in spatial patterns of smoking for lung cancer and its subtypes using sequentially adjusted models.



The significantly positive relative risks are highlighted by the red color and the DICs are listed for each model.

Figure 47: Relative risks and their 95% CIs per IQR increase for reconstructed smoking data (estimated spatial pattern+temporal pattern of smoking) for lung cancer and its subtypes using sequentially adjusted models.

In the regression models for O_3, we sequentially added the confounding effects that included the spatial random effect, the temporal fixed effect, and the spatial pattern of smoking, income, education and commuting for different health outcomes. We extracted the relative risks per IQR increase of O_3, their 95% CIs and DICs for the sequence of models as shown in Figure 45. According to the results for the sequentially adjusted models, O_3's effects on lung cancer and its subtypes are robust as they are not substantially influenced by other potential confounding effects. The results also suggest that exposure to O_3 is consistently and significantly associated with adenocarcinoma.

In the regression models for the spatial pattern for smoking, we constructed the sequentially adjusted models similar to the models for O_3 and the results are displayed in Figure 46. The results suggest that socioeconomic factors, especially income are influential confounding factors for smoking, particularly for total lung cancer, adenocarcinoma, and small cell carcinoma. We also did the analysis using the reconstructed smoking data (estimated spatial pattern+temporal pattern for smoking) as shown in Figure 39, which produced results similar to the models using the spatial patterns for smoking only.

The temporal trends estimated from the hierarchical models and the Poisson regressions are shown in Figure 48. The figures show generally increasing trends for the risks of lung cancer and all its subtypes except for large cell carcinoma from 2001 to 2007.

Finally, in order to interpret health effects for all the risk factors, we summarized their relative risks and their 95% CIs estimated from all the adjusted spatiotemporal models and independent Poisson models in Table 13. Comparing the two types of models, even though the relative risks estimated from the two models are consistent with each other, the spatiotemporal hierarchical models always outperform the independent Poisson models according to the DICs.

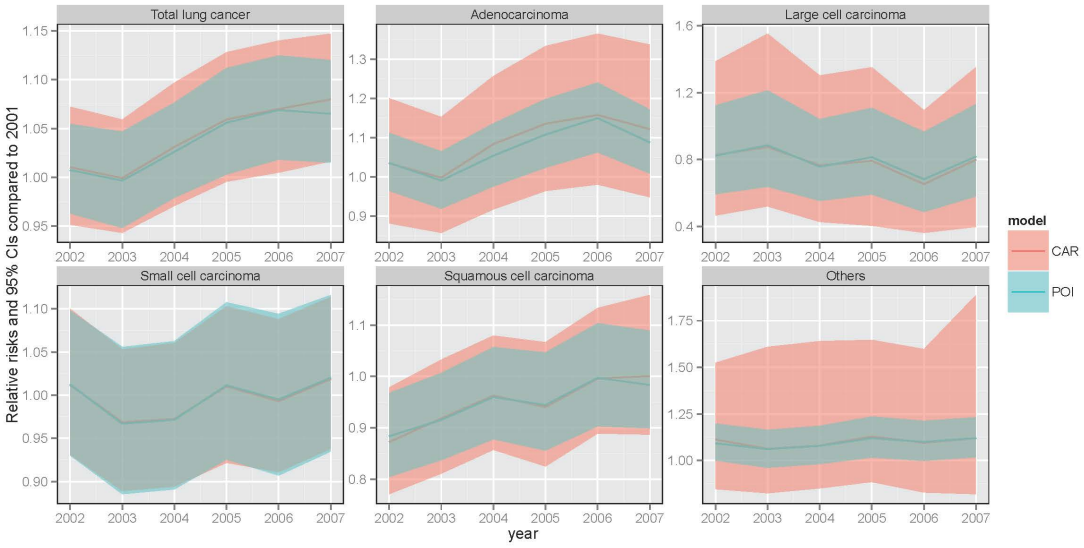


Figure 48: Relative risks and their 95% CIs for the years from 2002 to 2007 compared with the year 2001 for lung cancer and its subtypes.

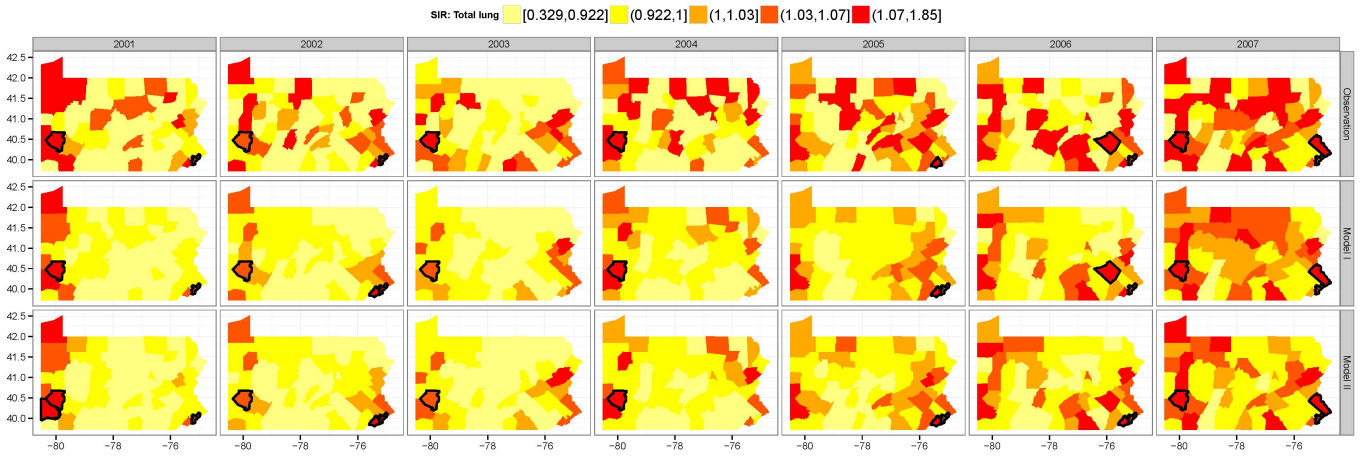
Table 13: Summary of relative risks per IQR increase in risk factors estimated from spatiotemporal hierarchical model and independent Poisson model.

Cancer	Risk factors	Spatiotemporal model				Independent Poisson model			
		RR	lower CI	upper CI	DIC	RR	lower CI	upper CI	DIC
Total lung	O_3	1.0200	1.0009	1.0401	3599.38	1.0239	1.0047	1.0428	3599.43
	Spatial patterns of smoking	1.0586	1.0321	1.0874		1.0687	1.0406	1.0975	
	Median household income	1.0371	1.0000	1.0723		1.0480	1.0115	1.0858	
	Percent of residences with education less than high school	0.9637	0.9413	0.9871		0.9674	0.9454	0.9895	
	Commuting time	1.0087	0.9919	1.0279		1.0020	0.9843	1.0185	
Adenocarcinoma	O_3	1.0495	1.0201	1.0788	2942.63	1.0615	1.0318	1.0912	2972.27
	Spatial patterns of smoking	1.0644	1.0224	1.1109		1.0630	1.0227	1.1062	
	Median household income	1.0701	1.0094	1.1332		1.0745	1.0243	1.1320	
	Percent of residences with education less than high school	0.9619	0.9252	0.9995		0.9633	0.9323	0.9976	
	Commuting time	0.9944	0.9659	1.0227		0.9883	0.9610	1.0154	
Large cell	O_3	1.0072	0.8859	1.1384	1689.12	1.0230	0.8990	1.1588	1690.19
	Spatial patterns of smoking	1.0687	0.8973	1.2680		1.0597	0.8896	1.2603	
	Median household income	0.9542	0.7446	1.2446		0.9229	0.7354	1.1571	
	Percent of residences with education less than high school	0.8831	0.7405	1.0705		0.8498	0.7278	0.9838	
	Commuting time	1.0258	0.9076	1.1594		0.9893	0.8821	1.1040	
Small cell	O_3	1.0140	0.9822	1.0466	2508.50	1.0139	0.9813	1.0490	2510.56
	Spatial patterns of smoking	1.0917	1.0392	1.1466		1.0943	1.0387	1.1514	
	Median household income	1.0200	0.9612	1.0881		1.0253	0.9592	1.0903	
	Percent of residences with education less than high school	0.9600	0.9247	0.9971		0.9600	0.9230	0.9986	
	Commuting time	1.0092	0.9758	1.0443		1.0082	0.9728	1.0422	
Squamous cell	O_3	1.0199	0.9845	1.0573	2740.13	1.0234	0.9885	1.0617	2746.47
	Spatial patterns of smoking	1.0538	1.0024	1.1103		1.0558	1.0048	1.1115	
	Median household income	0.9001	0.8385	0.9642		0.9039	0.8464	0.9652	
	Percent of residences with education less than high school	0.8742	0.8349	0.9163		0.8763	0.8397	0.9150	
	Commuting time	1.0438	1.0087	1.0817		1.0448	1.0087	1.0801	
Others	O_3	1.0118	0.9767	1.0487	3041.83	1.0039	0.9680	1.0427	3061.12
	Spatial patterns of smoking	1.0254	0.9712	1.0811		1.0743	1.0218	1.1298	
	Median household income	1.0690	0.9824	1.1546		1.1238	1.0524	1.1988	
	Percent of residences with education less than high school	1.0347	0.9857	1.0876		1.0352	0.9919	1.0793	
	Commuting time	1.0113	0.9798	1.0448		1.0067	0.9753	1.0390	

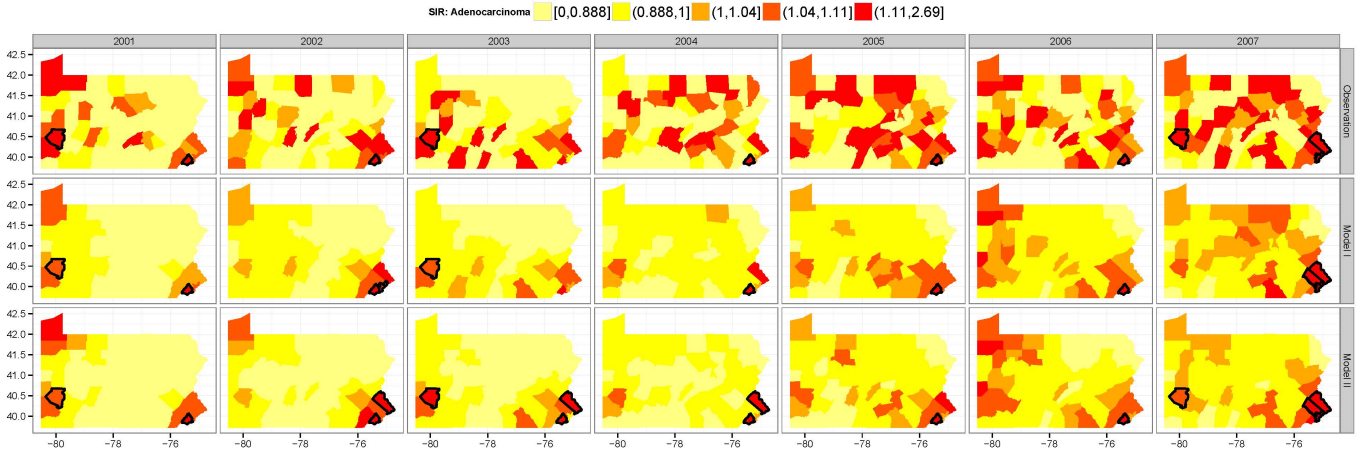
6.4 DISCUSSION

According to our statistical analysis, lung cancer incidence is increased by 2.00% per IQR increase in exposure to O_3, compared to 4.95% for adenocarcinoma, 0.72% for large cell, 1.40% for small cell, 1.90% for squamous cell carcinoma and 1.18% for other subtypes. Lung cancer incidence is increased by 5.86% per IQR increase in the spatial pattern of smoking, compared to 6.44% for adenocarcinoma, 6.87% for large cell, 9.17% for small cell, 5.38% for squamous cell carcinoma and 2.54% for other subtypes. Among the four major subtypes, O_3 is most associated with adenocarcinoma, while smoking is most associated with small cell carcinoma, which is consistent with the previous study [Kenfield et al., 2008]. In addition, our study also suggests that lung cancer risk induced by O_3 or usually comparable but lower than those induced by smoking.

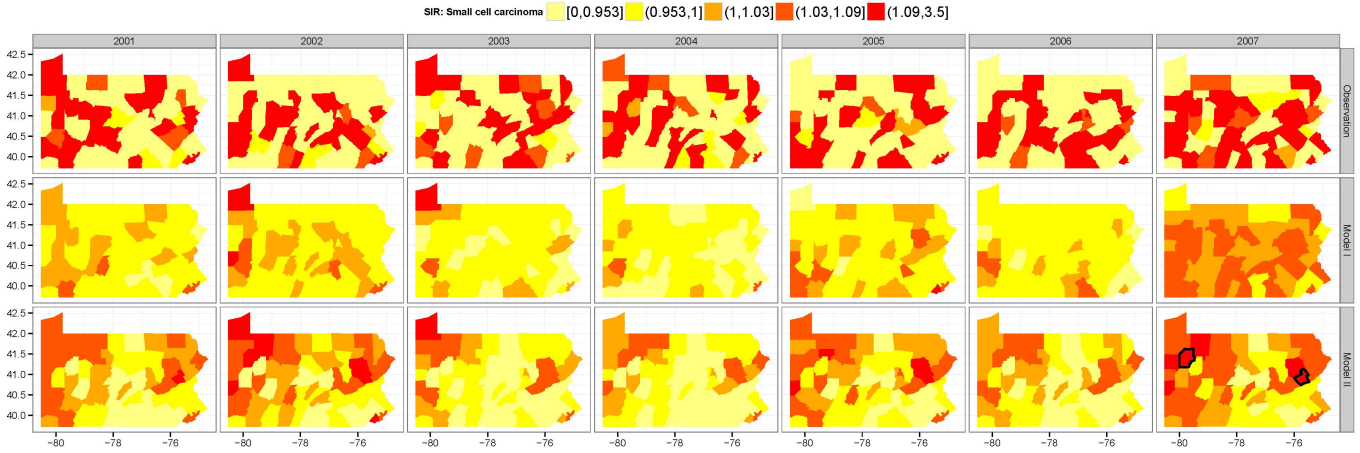
Among all the socioeconomic risk factors, a per IQR increase in median household income is significantly associated with a 3.71% increase in total lung cancer, a 7.01% increase in adenocarcinoma, and a 9.99% decrease in squamous cell carcinoma; a per IQR increase in percent of residences with education less than high school is significantly associated with a 3.63% decrease in total lung cancer, a 3.81% decrease in adenocarcinoma, a 4.00% decrease in small cell, and a 12.58% decrease in squamous cell carcinoma. Socioeconomic factors usually influence cancer risks in different respects. For example, diet, disease screening rate and lifestyle may play a role. Our analysis suggests that higher income and education levels are associated with higher lung cancer risks in Pennsylvania, which may be due to higher screening rates for wealthier populations.



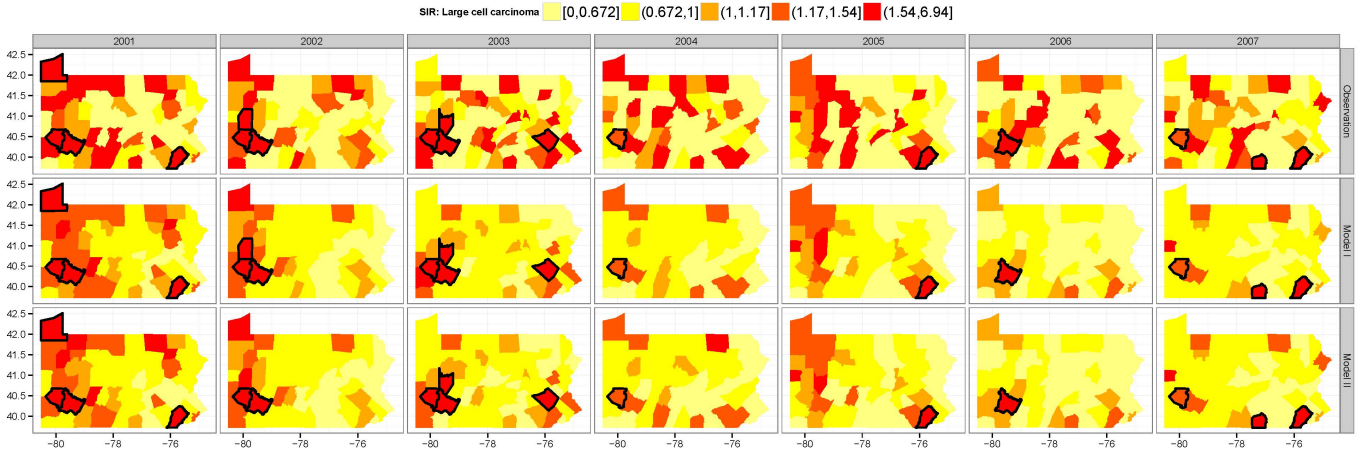
(a) Total lung cancer



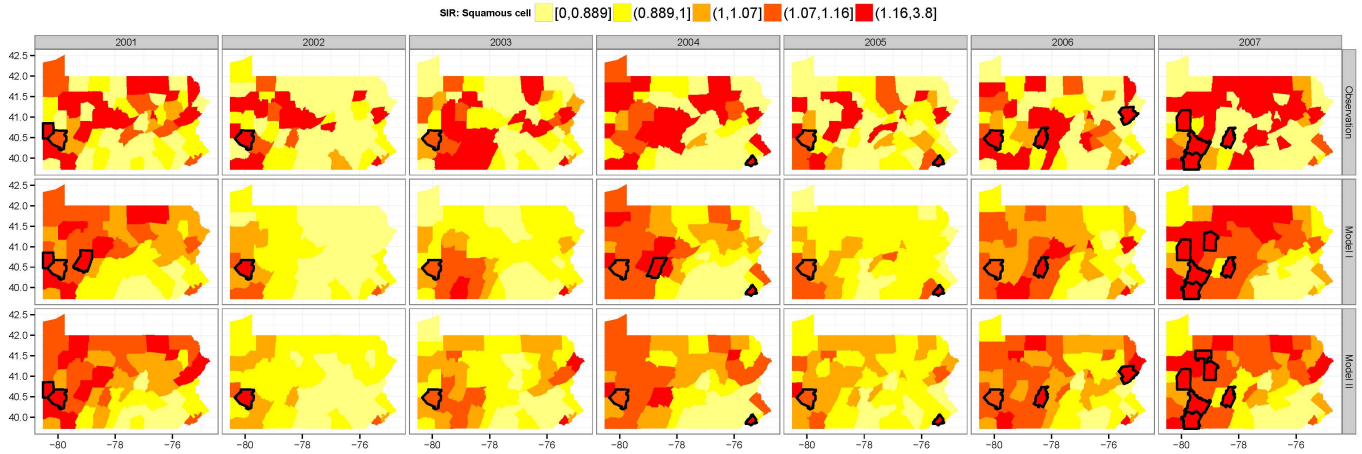
(b) Adenocarcinoma



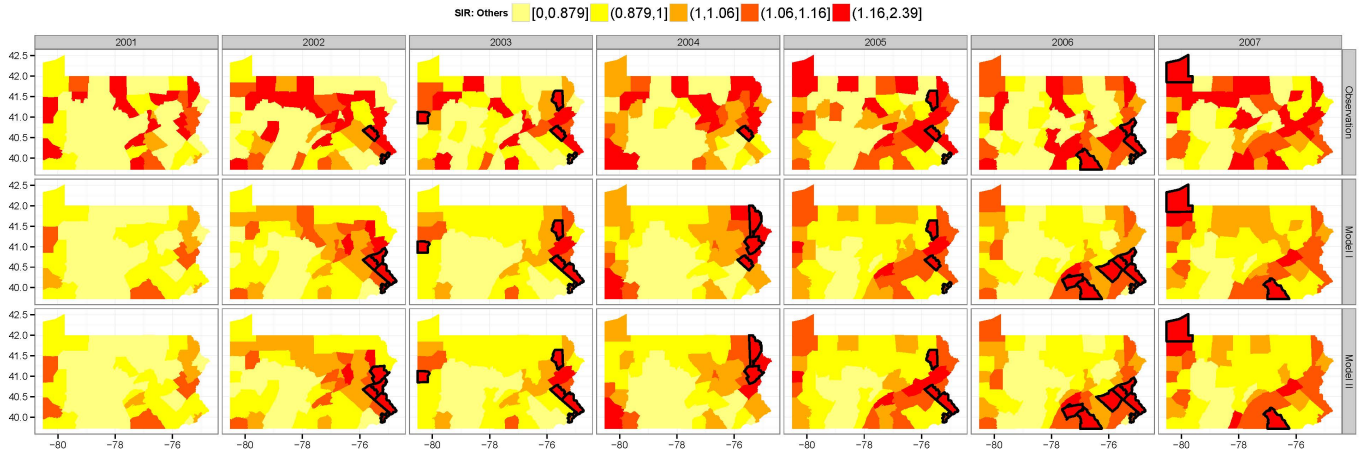
(c) Small cell carcinoma



(d) Large cell carcinoma



(e) Squamous cell carcinoma



(f) Others

Model I is a spatiotemporal smoothing model without any covariates and model II is a fully adjusted spatiotemporal hierarchical model. Counties with significantly higher risks are outlined in black polygons.

Figure 49: Comparison between observed SIRs, and model-fitted SIRs and identifying counties with significantly higher risks for lung cancer and its subtypes.

In addition, all counties with significantly higher risks to lung cancer and its subtypes can be identified by both observed SIRs and the model-fitted SIRs as shown in Figure 49 by comparing the lower boundaries of the 95% CI for observed or modeled counts and expected counts. To identify these risky counties, we displayed fitted SIRs for two types of models: a spatiotemporal smoothing model without any covariates and a fully adjusted spatiotemporal hierarchical model. Even though the fitted values of the two models are sightly different from each other, Allegheny County is always identified as a risky area consistently by the two models and observations for total lung cancer, adenocarcinoma, large cell and squamous cell carcinoma. However, our modeling results are limited in several respects. Firstly, the exposure data for O_3 that were generated by an interpolation model have an large uncertainty as shown by the 10-fold cross-validation (Figure 39) and may over-smooth O_3, as interpolated values were produced though averaging their spatiotemporal neighbors and the raw AQS data contains large numbers of missing values. The over-smoothness in interpolating results may cause potential exposure misclassification due to underestimation of the variance. Secondly, as historical records of county level smoking data are not available, we used the entire long-term trend of smoking in the US to represent smoking in the state of Pennsylvania and the spatial pattern of smoking to predict variations in cancer risks, which leads to similar regression coefficients for the spatial pattern of and reconstructed smoking data as shown in Figure 50. As we only captured the spatial pattern of smoking in our analysis, our models may underestimate its health effects on lung cancer and its subtypes. Thirdly, we obtained the socioeconomic factors and demographic information from the 2000 census, which may reflect the current population characteristics related to future lung cancer cases but not the population characteristics at the earlier exposure period. In addition, we did not control for other potential risk factors including radon radiation, diet and occupational exposure to hazardous chemicals, which may also bias our modeling results.

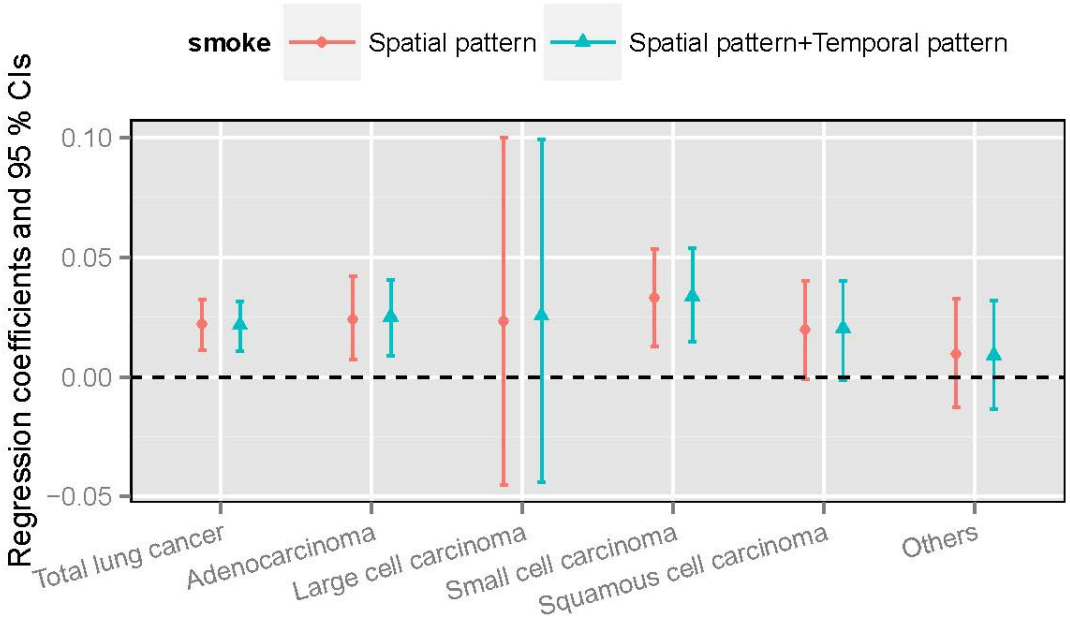


Figure 50: Comparing regression coefficients for the spatial pattern of smoking and reconstructed smoking data.

6.5 CONCLUSION

In this chapter, we explored the health effects of chronic exposure to O_3 on lung cancer incidences and its four major subtypes, including adenocarcinoma, large cell, small cell and squamous cell carcinoma, with adjustments of smoking, sex, age, race and socioeconomic factors. Even though our study is limited to an ecological study design and having poor quality historical data for O_3 and smoking, it finds that per IQR increase in O_3 is significantly associated with a 2.00% increase in the risk of lung cancer and a 4.95% increase in adenocarcinoma.

7.0 CONCLUSIONS

In this section, we will describe the public health significances of our studies in two respects: (1) findings in air pollution exposures and their health risks in the Pittsburgh region and (2) the use and their potential extensions of the statistical methods that have been developed in the above sections to environmental and epidemiological studies.

7.1 AIR POLLUTANTS AND THEIR HEALTH RISKS IN THE PITTSBURGH REGION FROM 2001 TO 2008

In the previously described studies, we have visualized time series of air pollutants (Figures 2 and 38), mortality counts (Figure 2), cancer incidences (Figures 6 and 34) and spatial patterns of air pollutants (Figures 27 and 42), standardized mortality ratios (Figure 26), standardized incidence ratios for cancer (Figures 7 and 36), some socioeconomic risk factors (Figures 8 and 37) in the Pittsburgh region (Allegheny County or the study domain of Allegheny, Washington and Westmoreland counties). These data and analyses will be useful for illustrating the public health issue of air pollution in the first ten years of twenty-first century in the Pittsburgh region in various respects which will be described in details in this section.

7.1.1 Temporal Trends of Air Pollutants and their Health Effects

In this paper, we explored both recent time-series of six air pollutants ($PM_{2.5}$, PM_{10}, NO_2, O_3, SO_2, CO) in the study domain of three counties from 1999 to 2009 (Figure 2) and an

historical time-series of O_3 in Allegheny County from 1980 to 2013 (Figure 38). According to the annual O_3 data, from 1980 to 2013 the pollutant decreased by about 50%. (Figure 38 shows that O_3 decreased from 0.12 ppm in 1980s to 0.06 ppm in 2010s in Allegheny County.) In addition, our daily time-series for $PM_{2.5}$, PM_{10} and O_3 (Figure 2) are also shown as the slightly decreased from 1999 to 2009 and more so for NO_2, SO_2, and CO. Even though the long-term decreasing trends of these time-series reflect the city of Pittsburgh's efforts in controlling air pollution, some of the daily pollutants were still above EPA criteria in recent years, especially for CO (8-hour criteria: 9 ppm) and $PM_{2.5}$ (24-hour criteria: 35 $\mu g/m^3$). However, the decreased levels of air pollutants may cause difficulties in estimating their acute health effects, because lower levels of pollutant might induce less temporal variations, which will increase estimated standard errors for the corresponding health effects. Previous studies have already associated air pollutants with mortality risk in the Pittsburgh region using time-series studies. For example, Mazumdar and Sussman, (1983) associated hourly measurements of particulate matter and SO_2 at three monitoring stations located at Hazelwood, Bellevue and Logans Ferry in Allegheny County with total mortality and mortality from heart disease for 1972 to 1977 and reported statistically significant health effects only for particulate matters measured by the monitor located at Hazelwood [Mazumdar and Sussman, 1983]. The study of Mazumdar and Sussman, (1983) showed that in the Pittsburgh region, time-series may fail to exhibit enough variation to detect their acute health effects, providing the impetus to design the spatiotemporal study in Chapter 5. In addition, Figure 38 shows that the decreasing trend of O_3 in the Pittsburgh region is similar to the trends in other counties, and indicates that spatial variation rather than temporal variation of air pollutants may be more critical for determining chronic health effects.

7.1.2 Spatial Trends of Air Pollutants and their Health Effects

In our studies, we expended efforts to explore and visualize spatial variations of air pollutants within the Pittsburgh region (Figure 27 and Table 10) and between Allegheny County and other counties in Pennsylvania (Figures 42 and 43) and spatial variations of mortality in the Pittsburgh region (Figure 26) and lung cancer incidences in the state of Pennsylvania

(Figures 36 and 44). Our analysis (Figure 27 and Table 10) shows that within the Pittsburgh region, air pollutants in the city of Pittsburgh were higher than the average level except for O_3 and PM_{10}, which coincides with the spatial patterns of mortalities (Figure 27). Another hot-spot for $PM_{2.5}$, PM_{10} and SO_2 within the Pittsburgh region was Liberty Borough. Among 67 counties in Pennsylvania, our studies show that Allegheny County had the fourth highest O_3 and the eighth highest lung cancer incidence, which indicated both air pollution and lung cancer are critical public health issues in the state of Pennsylvania. In Chapter 6, our statistical modeling results further confirmed the lung cancer issue in Allegheny County through identifying Allegheny County as one of the areas with significantly higher risks of lung cancer, adenocarcinoma, large cell carcinoma and small cell carcinoma than the average level among all the 67 counties of Pennsylvania. In addition, our studies show that the social disparity exists not only in the air pollutants (Figure 27), mortalities (Figure 26) and cancer incidences (Figure 7) but also in socioeconomic status (Figure 8) in the Pittsburgh region. Our analysis suggests that residents living in the city of Pittsburgh (especially the downtown area) usually had a lower SES index and higher percentage of smoking, air pollution exposure and mortality risks for cardiovascular and respiratory diseases.

7.2 COMMENTS ON THE STATISTICAL MODELS IN OUR STUDY

In our studies, we developed four spatiotemporal methods to model air pollutants and their health risks: (1) the two-step spatiotemporal regression model, (2) the hierarchical latent vector spatial model, (3) the spatiotemporal generalized estimating equations, and (4) the Bayesian hierarchical spatiotemporal model. Among the four models, the first two models estimated spatial or spatiotemporal variations of air pollutants by combining routine monitoring data with satellite measurements and the latter two models regressed their health outcomes on air pollutants while controlling spatiotemporal autocorrelations. In this section, we will discuss the statistical thinking behind the two types of models and the potential extensions of these models.

7.2.1 Statistical Methods for Combining Routine Monitoring and Satellite Measurements of Air Pollutants

As air pollutants can be observed by various methods, for example, routine monitors, atmospheric models (e.g. CMAQ) and satellite measurements, developing a statistical model to combine different types of measurements is an important problem to be solved in order to increase prediction accuracy for exposure assessment. In order to combine satellite measurements with monitoring data, we first developed a straight forward two-step model that first calibrated spatial misalignments between satellite measurements and locations of air monitors and then associated calibrated satellite measurements to monitoring data using a regression model. However, in the second step, using the calibrated satellite measurements as a covariate ignores the uncertainty in satellite measurements. Therefore, the two-step model evolved into a latent vector hierarchical model as described in Chapter 4. In the model we assumed a latent vector to represent the true values of air pollutant and modeled measurement errors of monitors and satellite instruments and spatial or spatiotemporal smoothness of the latent vector simultaneously. Our latent vector model suggests a general framework to combine various types of spatiotemporal measurements. Let $\boldsymbol{x}$ denote the latent vector, $\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(k)}$ denote k types of measurements of $\boldsymbol{x}$ and $\boldsymbol{Z}^{(i)}(i = 1, \cdots, k)$ denote the confoundors that influence i^{th} type measurements ($\boldsymbol{y}^{(i)}$) of $\boldsymbol{x}$ and then the general framework will include $k + 1$ parts, including k regression models: $\boldsymbol{y}^{(i)} = \boldsymbol{x}\beta^{(i)} + \boldsymbol{Z}^{(i)}\boldsymbol{b}^{(i)} + \boldsymbol{\epsilon}^{(i)}, i = 1, \cdots, k$ and another process model of $\boldsymbol{x} \sim \text{AR}(\cdot)$, $\boldsymbol{x} \sim \text{CAR}(\cdot)$ or $\boldsymbol{x} \sim \text{VAR}(\cdot)$ to consider the temporal, spatial or spatiotemporal autocorrelation of the latent vector $\boldsymbol{x}$. (For how to use a VAR process to model spatiotemporal autocorrelation, see page 383-384, Cressie and Wikle, 2011.) Model inference under such a framework can be performed generally by using the EM-algorithm. The framework can also be modified by introducing the well developed regression techniques such as ridge regression and Lasso into the k regression parts. The method can not only be applied to estimate air pollution but also other environmental data with multiple measurement methods.

7.2.2 Spatiotemporal Regression Methods

In Chapters 5 and 6, we applied two regression methods to associate air pollution to spatiotemporal health outcomes including the spatiotemporal generalized estimating equations and the Bayesian hierarchical model. In fact, the Bayesian hierarchical model has been widely used in previous regression analysis [Waller et al., 1997, Mariella and Tarantino, 2010], but may be not appropriate for a massive dataset due to the computational burden. Therefore, we developed the novel method of spatiotemporal generalized estimating equations for the regression analysis of massive daily mortality data. Actually, a similar spatiotemporal model has already been developed by previous researchers. For example, Cressie and Wikle, (2011) has proposed the dynamic spatiotemporal models (DSTM), in which the spatiotemporal autocorrelation can be modeled by a VAR process of a latent variable similar to our model specification in Section 5.2.1. However, in DSTM methodology, the latent spatiotemporal variable and all the parameters should be estimated together using the method of the Kalman filter and the EM-algorithm. For details of model inference for a DSTM, see pages 444-454, Cressie and Wikle, (2011). However, in most epidemiological studies, the latent spatiotemporal variable can be treated as a nuisance variable. Therefore, we extended Zeger's estimating equations for the regression coefficients in time-series analysis to spatiotemporal analysis [Zeger, 1988], which avoids computational burden. Even though our spatiotemporal generalized estimating equations are not as flexible as a DSTM, our method may be more practical for spatiotemporal regression of massive epidemiological datasets.

BIBLIOGRAPHY

[Adam Poupart et al., 2014] Adam Poupart, A., Brand, A., Fournier, M., Jerrett, M., and Smargiassi, A. (2014). Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of Kriging, Land-Use Regression (LUR), and combined Bayesian Maximum Entropy–LUR Approaches. *Environmental Health Perspectives*.

[Anderson et al., 1997] Anderson, H., Spix, C., Medina, S., Schouten, J., Castellsague, J., Rossi, G., Zmirou, D., Touloumi, G., Wojtyniak, B., Ponka, A., et al. (1997). Air pollution and daily admissions for chronic obstructive pulmonary disease in 6 European cities: results from the APHEA project. *European Respiratory Journal*, 10(5):1064–1071.

[Anselin, 1982] Anselin, L. (1982). A note on small sample properties of estimators in a first-order spatial autoregressive model. *Environment and Planning A*, 14(8):1023–30.

[Anselin, 1995] Anselin, L. (1995). Local indicators of spatial association–LISA. *Geographical analysis*, 27(2):93–115.

[Banerjee et al., 2004] Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical modeling and analysis for spatial data*. CRC Press.

[Banerjee et al., 2003] Banerjee, S., Wall, M. M., and Carlin, B. P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, 4(1):123–142.

[Beelen et al., 2009] Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., and Briggs, D. J. (2009). Mapping of background air pollution at a fine spatial scale across the European Union. *Science of the Total Environment*, 407(6):1852–1867.

[Beeson et al., 1998] Beeson, W. L., Abbey, D. E., and Knutsen, S. F. (1998). Long-term concentrations of ambient air pollutants and incident lung cancer in California adults: results from the AHSMOG study. Adventist Health Study on Smog. *Environmental Health Perspectives*, 106(12):813.

[Bernhard Pfaff, 2008] Bernhard Pfaff (2008). VAR, SVAR and SVEC Models: Implementation Within R Package vars. *Journal of Statistical Software*, 27(4).

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

[Bilonick et al., 2015] Bilonick, R. A., Connell, D. P., Talbott, E. O., Rager, J. R., and Xue, T. (2015). Using structural equation modeling to construct calibration equations relating PM2.5 mass concentration samplers to the federal reference method sampler. *Atmospheric Environment*, 103:365–377.

[Boos and Stefanski, 2013] Boos, D. D. and Stefanski, L. (2013). *Essential Statistical Inference*. Springer.

[Brauer et al., 2002] Brauer, M., Hoek, G., Van Vliet, P., Meliefste, K., Fischer, P. H., Wijga, A., Koopman, L. P., Neijens, H. J., Gerritsen, J., Kerkhof, M., et al. (2002). Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children. *American Journal of Respiratory and Critical Care Medicine*, 166(8):1092–1098.

[Brook et al., 2004] Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., Smith, S. C., et al. (2004). Air pollution and cardiovascular disease A statement for healthcare professionals from the expert panel on population and prevention science of the American Heart Association. *Circulation*, 109(21):2655–2671.

[Brunekreef et al., 1997] Brunekreef, B., Janssen, N. A., de Hartog, J., Harssema, H., Knape, M., and van Vliet, P. (1997). Air pollution from truck traffic and lung function in children living near motorways. *Epidemiology*, pages 298–303.

[Bruno et al., 2009] Bruno, F., Guttorp, P., Sampson, P. D., and Cocchi, D. (2009). A simple non-separable, non-stationary spatiotemporal model for ozone. *Environmental and Ecological Statistics*, 16(4):515–529.

[Chen et al.,] Chen, C.-L., Hsu, L.-I., Chiou, H.-Y., Hsueh, Y.-M., Chen, S.-Y., Wu, M.-M., Chen, C.-J., Blackfoot Disease Study Group, et al. Ingested arsenic, cigarette smoking, and lung cancer risk: a follow-up study in arseniasis-endemic areas in Taiwan. *JAMA: the Journal of the American Medical Association*.

[Cheng et al., 2009] Cheng, I., Witte, J. S., McClure, L. A., Shema, S. J., Cockburn, M. G., John, E. M., and Clarke, C. A. (2009). Socioeconomic status and prostate cancer incidence and mortality rates among the diverse population of California. *Cancer Causes & Control*, 20(8):1431–1440.

[Choi et al., 2009] Choi, J., Reich, B. J., Fuentes, M., and Davis, J. M. (2009). Multivariate spatial-temporal modeling and prediction of speciated fine particles. *Journal of Statistical Theory and Practice*, 3(2):407–418.

[Cleveland, 1979] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

[Cliff and Ord, 1981] Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models & applications*, volume 44. Pion London.

[Clougherty et al., 2008] Clougherty, J. E., Wright, R. J., Baxter, L. K., Levy, J. I., et al. (2008). Land use regression modeling of intra-urban residential variability in multiple traffic-related air pollutants. *Environmental Health*, 7(1):17.

[Cohan et al., 2005] Cohan, D. S., Hakami, A., Hu, Y., and Russell, A. G. (2005). Nonlinear response of ozone to emissions: Source apportionment and sensitivity analysis. *Environmental Science & Technology*, 39(17):6739–6748.

[Correia et al., 2013] Correia, A. W., Pope III, C. A., Dockery, D. W., Wang, Y., Ezzati, M., and Dominici, F. (2013). The effect of air pollution control on life expectancy in the United States: an analysis of 545 US counties for the period 2000 to 2007. *Epidemiology (Cambridge, Mass.)*, 24(1):23.

[Coyle et al., 2006] Coyle, Y. M., Minahjuddin, A. T., Hynan, L. S., and Minna, J. D. (2006). An ecological study of the association of metal air pollutants with lung cancer incidence in Texas. *Journal of Thoracic Oncology*, 1(7):654–661.

[Cressie, 1993] Cressie, N. (1993). Statistics for spatial data.

[Cressie and Huang, 1999] Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339.

[Cressie and Majure, 1997] Cressie, N. and Majure, J. J. (1997). Spatio-temporal statistical modeling of livestock waste in streams. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 24–47.

[Cressie and Wikle, 2011] Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

[Crumeyrolle et al., 2013] Crumeyrolle, S., Chen, G., Ziemba, L., Beyersdorf, A., Thornhill, L., Winstead, E., Moore, R., Shook, M., and Anderson, B. (2013). Factors that influence surface PM 2.5 values inferred from satellite observations: perspective gained for the Baltimore-Washington Area during DISCOVER-AQ. *Atmospheric Chemistry and Physics Discussions*, 13(9):23421–23459.

[CVX Research, 2012] CVX Research, I. (2012). CVX: Matlab software for disciplined convex programming, version 2.0. http://cvxr.com/cvx.

[Darby et al., 2005] Darby, S., Hill, D., Auvinen, A., Barros Dios, J., Baysson, H., Bochicchio, F., Deo, H., Falk, R., Forastiere, F., Hakama, M., et al. (2005). Radon in homes and

risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *BMJ*, 330(7485):223.

[Davis et al., 2012] Davis, R. A., Zang, P., and Zheng, T. (2012). Sparse vector autoregressive modeling. *arXiv preprint arXiv:1207.0520*.

[De Iaco, 2010] De Iaco, S. (2010). Space–time correlation analysis: a comparative study. *Journal of Applied Statistics*, 37(6):1027–1041.

[De Iaco et al., 2002a] De Iaco, S., Myers, D., and Posa, D. (2002a). Space–time variograms and a functional form for total air pollution measurements. *Computational statistics & data analysis*, 41(2):311–328.

[De Iaco et al., 2002b] De Iaco, S., Myers, D. E., and Posa, D. (2002b). Nonseparable space-time covariance models: some parametric families. *Mathematical Geology*, 34(1):23–42.

[Devlin et al., 1991] Devlin, R. B., McDonnell, W. F., Mann, R., Becker, S., House, D. E., Schreinemachers, D., and Koren, H. S. (1991). Exposure of humans to ambient levels of ozone for 6.6 hours causes cellular and biochemical changes in the lung. *American Journal of Respiratory Cell and Molecular Biology*, 4(1):72–81.

[Di Giacinto, 2010] Di Giacinto, V. (2010). On vector autoregressive modeling in space and time. *Journal of Geographical Systems*, 12(2):125–154.

[Diggle et al., 2002] Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

[Dijkema et al., 2011] Dijkema, M., Mallant, S. F., Gehring, U., van den Hurk, K., Alssema, M., van Strien, R. T., Fischer, P. H., Nijpels, G., Stehouwer, C. D., Hoek, G., et al. (2011). Long-term exposure to traffic-related air pollution and type 2 diabetes prevalence in a cross-sectional screening-study in the Netherlands. *Environmental Health*, 10:76.

[Dockery et al., 1989] Dockery, D. W., Speizer, F. E., Stram, D. O., Ware, J. H., Spengler, J. D., and Ferris Jr, B. G. (1989). Effects of inhalable particles on respiratory health of children. *American Review of Respiratory Disease*, 139(3):587–594.

[Dockery et al., 1982] Dockery, D. W., Ware, J. H., Ferris Jr, B. G., Speizer, F. E., Cook, N. R., and Herman, S. M. (1982). Change in pulmonary function in children associated with air pollution episodes. *Journal of the Air Pollution Control Association*, 32(9):937–942.

[Doll, 1993] Doll, R. (1993). Mortality from lung cancer in asbestos workers 1955. *British Journal of Industrial Medicine*, 50(6):485.

[Dominici et al., 2006] Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA: the Journal of the American Medical Association*, 295(10):1127–1134.

[Dueñas et al., 2002] Dueñas, C., Fernández, M., Cañete, S., Carretero, J., and Liger, E. (2002). Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. *Science of the Total Environment*, 299(1):97–113.

[Dwyer Lindgren et al., 2014] Dwyer Lindgren, L., Mokdad, A. H., Srebotnjak, T., Flaxman, A. D., Hansen, G. M., and Murray, C. J. (2014). Cigarette smoking prevalence in US counties: 1996-2012. *Population Health Metrics*, 12(1):5.

[Eder and Yu, 2006] Eder, B. and Yu, S. (2006). A performance evaluation of the 2004 release of Models-3 CMAQ. *Atmospheric Environment*, 40(26):4811–4824.

[Edzer J. Pebesma, 2004] Edzer J. Pebesma (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691.

[Engel Cox et al., 2004] Engel Cox, J. A., Holloman, C. H., Coutant, B. W., and Hoff, R. M. (2004). Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. *Atmospheric Environment*, 38(16):2495–2509.

[Frischer et al., 1999] Frischer, T., Studnicka, M., GARTNER, C., Tauber, E., Horak, F., Veiter, A., Spengler, J., Kuhr, J., and Urbanek, R. (1999). Lung function growth and ambient ozone: a three-year population study in school children. *American Journal of Respiratory and Critical Care Medicine*, 160(2):390–396.

[Fuentes, 2001] Fuentes, M. (2001). A high frequency Kriging approach for non-stationary environmental processes. *Environmetrics*, 12(5):469–483.

[Fuentes and Raftery, 2005] Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45.

[Gamble, 1998] Gamble, J. F. (1998). PM2.5 and mortality in long-term prospective cohort studies: cause-effect or statistical associations? *Environmental Health Perspectives*, 106(9):535.

[Gan et al., 2013] Gan, W. Q., FitzGerald, J. M., Carlsten, C., Sadatsafavi, M., and Brauer, M. (2013). Associations of ambient air pollution with chronic obstructive pulmonary disease hospitalization and mortality. *American Journal of Respiratory and Critical Care Medicine*, 187(7):721–727.

[Gent et al.,] Gent, J. F., Triche, E. W., Holford, T. R., Belanger, K., Bracken, M. B., Beckett, W. S., and Leaderer, B. P. Association of low-level ozone and fine particles with respiratory symptoms in children with asthma. *JAMA: the Journal of the American Medical Association*.

[Genton, 2007] Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics*, 18(7):681–695.

[Glad et al., 2012] Glad, J. A., Brink, L. L., Talbott, E. O., Lee, P. C., Xu, X., Saul, M., and Rager, J. (2012). The Relationship of Ambient Ozone and PM2.5 Levels and Asthma Emergency Department Visits: Possible Influence of Gender and Ethnicity. *Archives of Environmental & Occupational Health*, 67(2):103–108.

[Gneiting, 2002] Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600.

[Guttorp et al., 1994] Guttorp, P., Meiring, W., and Sampson, P. D. (1994). A space-time analysis of ground-level ozone data. *Environmetrics*, 5(3):241–254.

[Hastie and Tibshirani,] Hastie, T. and Tibshirani, R. Generalized additive models.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

[Henderson et al., 2007] Henderson, S. B., Beckerman, B., Jerrett, M., and Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental Science & Technology*, 41(7):2422–2428.

[Hengl et al., 2004] Hengl, T., Heuvelink, G., and Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-Kriging. *Geoderma*, 120(1):75–93.

[Hoek et al., 2008] Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561–7578.

[Hoek et al., 2002] Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., and van den Brandt, P. A. (2002). Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *The lancet*, 360(9341):1203–1209.

[Iaco et al., 2001] Iaco, S. D., Myers, D. E., and Posa, D. (2001). Space–time analysis using a general product–sum model. *Statistics & Probability Letters*, 52(1):21–28.

[Ito et al., 2005] Ito, K., De Leon, S. F., and Lippmann, M. (2005). Associations between ozone and daily mortality: analysis and meta-analysis. *Epidemiology*, 16(4):446–457.

[Jaffe, 2010] Jaffe, D. (2010). Relationship between surface and free tropospheric ozone in the western US. *Environmental Science & Technology*, 45(2):432–438.

[Jerrett et al., 2005] Jerrett, M., Burnett, R. T., Ma, R., Pope III, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E., et al. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*, 16(6):727–736.

[Jerrett et al., 2009] Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E., and Thun, M. (2009). Long-term ozone exposure and mortality. *New England Journal of Medicine*, 360(11):1085–1095.

[Jin et al., 2005] Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, 61(4):950–961.

[Johansen, 1995] Johansen, S. (1995). Likelihood-based inference in cointegrated vector autoregressive models. *OUP Catalogue*.

[Jørgensen et al., 1996] Jørgensen, B., Lundbye Christensen, S., Song, P. X.-K., and Sun, L. (1996). State-space models for multivariate longitudinal data of mixed types. *Canadian Journal of Statistics*, 24(3):385–402.

[Jun and Stein, 2004] Jun, M. and Stein, M. L. (2004). Statistical comparison of observed and CMAQ modeled daily sulfate levels. *Atmospheric Environment*, 38(27):4427–4436.

[Katzfuss and Cressie, 2011] Katzfuss, M. and Cressie, N. (2011). Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32(4):430–446.

[Kenfield et al., 2008] Kenfield, S. A., Wei, E. K., Stampfer, M. J., Rosner, B. A., and Colditz, G. A. (2008). Comparison of aspects of smoking among the four histological types of lung cancer. *Tobacco Control*, 17(3):198–204.

[Kilian, 2011] Kilian, L. (2011). *Structural vector autoregressions*. Centre for Economic Policy Research.

[Kim et al., 2005] Kim, E., Hopke, P. K., Pinto, J. P., and Wilson, W. E. (2005). Spatial variability of fine particle mass, components, and source contributions during the regional air pollution study in St. Louis. *Environmental Science & Technology*, 39(11):4172–4179.

[Kumar et al., 2007] Kumar, N., Chu, A., and Foster, A. (2007). An empirical relationship between PM¡ sub¿ 2.5¡/sub¿ and aerosol optical depth in Delhi Metropolitan. *Atmospheric Environment*, 41(21):4492–4503.

[Laden et al., 2006] Laden, F., Schwartz, J., Speizer, F. E., and Dockery, D. W. (2006). Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. *American Journal of Respiratory and Critical Care Medicine*, 173(6):667–672.

[Lee, 2013] Lee, D. (2013). CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors. *Journal of Statistical Software*, 55(13):1–24.

[Levelt et al., 2006] Levelt, P. F., van den Oord, G. H., Dobber, M. R., Malkki, A., Visser, H., de Vries, J., Stammes, P., Lundell, J. O., and Saari, H. (2006). The ozone monitoring instrument. *Geoscience and Remote Sensing, IEEE Transactions on*, 44(5):1093–1101.

[Levy et al., 2001] Levy, D., Lumley, T., Sheppard, L., Kaufman, J., and Checkoway, H. (2001). Referent selection in case-crossover analyses of acute health effects of air pollution. *Epidemiology*, 12(2):186–192.

[Lichstein et al., 2002] Lichstein, J. W., Simons, T. R., Shriner, S. A., and Franzreb, K. E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, 72(3):445–463.

[Lin et al., 2011] Lin, W., Huang, W., Zhu, T., Hu, M., Brunekreef, B., Zhang, Y., Liu, X., Cheng, H., Gehring, U., Li, C., et al. (2011). Acute respiratory inflammation in children and black carbon in ambient air before and during the 2008 Beijing Olympics. *Environmental Health Perspectives*, 119(10):1507.

[Lindgren et al., 2009] Lindgren, A., Stroh, E., Montnémery, P., Nihlén, U., Jakobsson, K., and Axmon, A. (2009). Traffic-related air pollution associated with prevalence of asthma and COPD/chronic bronchitis. A cross-sectional study in Southern Sweden. *International Journal of Health Geographics*, 8(1):2.

[Liu et al.,] Liu, X.-H., Zhang, Y., Cheng, S.-H., Xing, J., Zhang, Q., Streets, D. G., Jang, C., Wang, W.-X., and Hao, J.-M. Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation. *Atmospheric Environment*, 44(20):2415–2426.

[Liu et al., 2007] Liu, Y., Franklin, M., Kahn, R., and Koutrakis, P. (2007). Using aerosol optical thickness to predict ground-level PM¡ sub¿ 2.5¡/sub¿ concentrations in the St. Louis area: A comparison between MISR and MODIS. *Remote Sensing of Environment*, 107(1):33–44.

[Liu et al., 2009] Liu, Y., Paciorek, C. J., and Koutrakis, P. (2009). Estimating regional spatial and temporal variability of PM2.5 concentrations using satellite data, meteorology, and land use information. *Environmental Health Perspectives*, 117(6):886.

[Loomis et al., 1996] Loomis, D. P., Borja Aburto, V., Bangdiwala, S., and Shy, C. (1996). Ozone exposure and daily mortality in Mexico City: a time-series analysis. *Research report (Health Effects Institute)*, (75):1–37.

[Lunn et al., 2000] Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

[Lütkepohl, 2006] Lütkepohl, H. (2006). *Structural vector autoregressive analysis for cointegrated variables*. Springer.

[Ma, 2003] Ma, C. (2003). Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference*, 116(2):489–501.

[Mao et al., 2001] Mao, Y., Hu, J., Ugnat, A.-M., Semenciw, R., Fincham, S., et al. (2001). Socioeconomic status and lung cancer risk in Canada. *International Journal of Epidemiology*, 30(4):809–817.

[Mariella and Tarantino, 2010] Mariella, L. and Tarantino, M. (2010). Spatial temporal conditional auto-regressive model: A new autoregressive matrix. *Australian J Stat*, 39(3):223.

[Martin, 2008] Martin, R. V. (2008). Satellite remote sensing of surface air quality. *Atmospheric Environment*, 42(34):7823–7843.

[Mazumdar and Sussman, 1983] Mazumdar, S. and Sussman, N. (1983). Relationships of air pollution to health: Results from the Pittsburgh Study. *Archives of Environmental Health: An International Journal*, 38(1):17–24.

[McConnell et al., 2002] McConnell, R., Berhane, K., Gilliland, F., London, S. J., Islam, T., Gauderman, W. J., Avol, E., Margolis, H. G., and Peters, J. M. (2002). Asthma in exercising children exposed to ozone: a cohort study. *The Lancet*, 359(9304):386–391.

[McDonnell et al., 1999] McDonnell, W. F., Abbey, D. E., Nishino, N., and Lebowitz, M. D. (1999). Long-term ambient ozone concentration and the incidence of asthma in nonsmoking adults: the AHSMOG Study. *Environmental Research*, 80(2):110–121.

[McMillan et al., 2010] McMillan, N. J., Holland, D. M., Morara, M., and Feng, J. (2010). Combining numerical model output and particulate data using Bayesian space–time modeling. *Environmetrics*, 21(1):48–65.

[Medina Ramón et al., 2006] Medina Ramón, M., Zanobetti, A., and Schwartz, J. (2006). The effect of ozone and PM10 on hospital admissions for pneumonia and chronic obstructive pulmonary disease: a national multicity study. *American Journal of Epidemiology*, 163(6):579–588.

[Menzies and Chahine, 1974] Menzies, R. and Chahine, M. (1974). Remote atmospheric sensing with an airborne laser absorption spectrometer. *Applied Optics*, 13(12):2840–2849.

[Mercer et al., 2011] Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., Avol, E. L., Oron, A. P., Larson, T., Liu, L.-J. S., et al. (2011). Comparing universal Kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmospheric Environment*, 45(26):4412–4420.

[Miller et al., 2007] Miller, K. A., Siscovick, D. S., Sheppard, L., Shepherd, K., Sullivan, J. H., Anderson, G. L., and Kaufman, J. D. (2007). Long-term exposure to air pollution and incidence of cardiovascular events in women. *New England Journal of Medicine*, 356(5):447–458.

[Monn, 2001] Monn, C. (2001). Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. *Atmospheric Environment*, 35(1):1–32.

[Mugglin et al., 2002] Mugglin, A. S., Cressie, N., and Gemmell, I. (2002). Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine*, 21(18):2703–2721.

[Nyberg et al., 2000] Nyberg, F., Gustavsson, P., Järup, L., Bellander, T., Berglind, N., Jakobsson, R., and Pershagen, G. (2000). Urban air pollution and lung cancer in Stockholm. *Epidemiology*, 11(5):487–495.

[Öberg et al., 2011] Öberg, M., Jaakkola, M. S., Woodward, A., Peruga, A., and Prüss Ustün, A. (2011). Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from 192 countries. *The Lancet*, 377(9760):139–146.

[Paciorek and Liu, 2008] Paciorek, C. J. and Liu, Y. (2008). Limitations of remotely-sensed aerosol as a spatial proxy for fine particulate matter.

[Paciorek et al., 2008] Paciorek, C. J., Liu, Y., Moreno-Macias, H., and Kondragunta, S. (2008). Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM2.5. *Environmental Science & Technology*, 42(15):5800–5806.

[Pawitan, 2001] Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

[Peng et al., 2009] Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2009). Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*, 117(6):957.

[Peters et al., 2001] Peters, A., Dockery, D. W., Muller, J. E., and Mittleman, M. A. (2001). Increased particulate air pollution and the triggering of myocardial infarction. *Circulation*, 103(23):2810–2815.

[Peto et al., 2000] Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., and Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ*, 321(7257):323–329.

[Pfaff, 2008] Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. Springer, New York, Second edition. ISBN 0-387-27960-1.

[Phillips et al., 1997] Phillips, D. L., Lee, E. H., Herstrom, A. A., Hogsett, W. E., and Tingey, D. T. (1997). Use of auxiliary data for spatial interpolation of ozone exposure in southeastern forests. *Environmetrics*, 8(1):43–61.

[Pinto et al., 2004] Pinto, J. P., Lefohn, A. S., and Shadwick, D. S. (2004). Spatial variability of PM2.5 in urban areas in the United States. *Journal of the Air & Waste Management Association*, 54(4):440–449.

[Pope et al., 1996] Pope, C. A. et al. (1996). Synoptic weather modeling and estimates of the exposure-response relationship between daily mortality and particulate air pollution. *Environmental Health Perspectives*, 104(4):414.

[Pope et al., 2006] Pope, C. A., Muhlestein, J. B., May, H. T., Renlund, D. G., Anderson, J. L., and Horne, B. D. (2006). Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution. *Circulation*, 114(23):2443–2448.

[Pope et al., 2011] Pope, C. r., Burnett, R. T., Turner, M. C., Cohen, A., Krewski, D., Jerrett, M., Gapstur, S. M., and Thun, M. J. (2011). Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure-response relationships. *Environmental Health Perspectives*, 119(11):1616–1621.

[Pope 3rd, 1989] Pope 3rd, C. (1989). Respiratory disease associated with community air pollution and a steel mill, Utah Valley. *American Journal of Public Health*, 79(5):623–628.

[Pope 3rd et al., 1999] Pope 3rd, C., Hill, R. W., and Villegas, G. M. (1999). Particulate air pollution and daily mortality on Utah's Wasatch Front. *Environmental Health Perspectives*, 107(7):567.

[Pope 3rd et al., 2004] Pope 3rd, C. A., Hansen, M. L., Long, R. W., Nielsen, K. R., Eatough, N. L., Wilson, W. E., and Eatough, D. J. (2004). Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. *Environmental Health Perspectives*, 112(3):339.

[Pope III et al.,] Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA: the Journal of the American Medical Association*.

[Pope III and Dockery, 1992] Pope III, C. A. and Dockery, D. W. (1992). Acute health effects of PM10 pollution on symptomatic and asymptomatic children. *American Review of Respiratory Disease*, 145(5):1123–1128.

[Pope III et al., 1991] Pope III, C. A., Dockery, D. W., Spengler, J. D., and Raizenne, M. E. (1991). Respiratory health and PM10 pollution: a daily time series analysis. *American Review of Respiratory Disease*, 144(3_pt_1):668–674.

[Pope III et al., 1992] Pope III, C. A., Schwartz, J., and Ransom, M. R. (1992). Daily mortality and PM10 pollution in Utah Valley. *Archives of Environmental Health: An International Journal*, 47(3):211–217.

[Pope III et al., 1995] Pope III, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., and Heath Jr, C. W. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American Journal of Respiratory and Critical Care Medicine*, 151(3_pt_1):669–674.

[Pudasainee et al., 2006] Pudasainee, D., Sapkota, B., Shrestha, M. L., Kaga, A., Kondo, A., and Inoue, Y. (2006). Ground level ozone concentrations and its association with NOx and meteorological parameters in Kathmandu valley, Nepal. *Atmospheric Environment*, 40(40):8081–8087.

[Reilly and Gelman, 2007] Reilly, C. and Gelman, A. (2007). Weighted classical variogram estimation for data with clustering. *Technometrics*, 49(2).

[Ren et al., 2010] Ren, C., Park, S. K., Vokonas, P. S., Sparrow, D., Wilker, E., Baccarelli, A., Suh, H. H., Tucker, K. L., Wright, R. O., and Schwartz, J. (2010). Air pollution and homocysteine: more evidence that oxidative stress-related genes modify effects of particulate air pollution. *Epidemiology (Cambridge, Mass.)*, 21(2):198.

[Rich et al., 2012] Rich, D. Q., Kipen, H. M., Huang, W., Wang, G., Wang, Y., Zhu, P., Ohman Strickland, P., Hu, M., Philipp, C., Diehl, S. R., et al. (2012). Association between changes in air pollution levels during the Beijing Olympics and biomarkers of inflammation and thrombosis in healthy young adults. *JAMA: the Journal of the American Medical Association*, 307(19):2068–2078.

[Richter et al., 2005] Richter, A., Burrows, J. P., Nüß, H., Granier, C., and Niemeier, U. (2005). Increase in tropospheric nitrogen dioxide over China observed from space. *Nature*, 437(7055):129–132.

[Ritz et al., 2000] Ritz, B., Yu, F., Chapa, G., and Fruin, S. (2000). Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993. *Epidemiology*, 11(5):502–511.

[Ritz et al., 2002] Ritz, B., Yu, F., Fruin, S., Chapa, G., Shaw, G. M., and Harris, J. A. (2002). Ambient air pollution and risk of birth defects in Southern California. *American Journal of Epidemiology*, 155(1):17–25.

[Robert et al., 2004] Robert, S. A., Trentham Dietz, A., Hampton, J. M., McElroy, J. A., Newcomb, P. A., and Remington, P. L. (2004). Socioeconomic risk factors for breast cancer: distinguishing individual-and community-level effects. *Epidemiology*, 15(4):442–450.

[Rodgers et al., 2000] Rodgers, C. D. et al. (2000). *Inverse methods for atmospheric sounding: theory and practice*, volume 2. World scientific Singapore.

[Roemer et al., 1993] Roemer, W., Hoek, G., and Brunekreef, B. (1993). Effect of ambient winter air pollution on respiratory health of children with chronic respiratory symptoms. *American Review of Respiratory Disease*, 147(1):118–124.

[Roger Bivand, 2014] Roger Bivand (2014). *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-74.

[Sahu et al., 2006] Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):61–86.

[Sahu et al., 2007] Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007). High-resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical Association*, 102(480):1221–1234.

[Salam et al., 2005] Salam, M. T., Millstein, J., Li, Y.-F., Lurmann, F. W., Margolis, H. G., and Gilliland, F. D. (2005). Birth outcomes and prenatal exposure to ozone, carbon monoxide, and particulate matter: results from the Children's Health Study. *Environmental Health Perspectives*, pages 1638–1644.

[Sang and Huang, 2012] Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

[Sapkota et al., 2012] Sapkota, A., Chelikowsky, A. P., Nachman, K. E., Cohen, A. J., and Ritz, B. (2012). Exposure to particulate matter and adverse birth outcomes: a comprehensive review and meta-analysis. *Air Quality, Atmosphere & Health*, 5(4):369–381.

[Schmid and Held, 2004] Schmid, V. and Held, L. (2004). Bayesian extrapolation of space–time trends in cancer registry data. *Biometrics*, 60(4):1034–1042.

[Schmidt and O'Hagan, 2003] Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.

[Schwartz, 1996] Schwartz, J. (1996). Air pollution and hospital admissions for respiratory disease. *Epidemiology*, pages 20–28.

[Schwartz, 1999] Schwartz, J. (1999). Air pollution and hospital admissions for heart disease in eight US counties. *Epidemiology*, 10(1):17–22.

[Schwartz and Morris, 1995] Schwartz, J. and Morris, R. (1995). Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan. *American Journal of Epidemiology*, 142(1):23–35.

[Sims, 1980] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

[Smith et al., 1998] Smith, A. H., Goycolea, M., Haque, R., and Biggs, M. L. (1998). Marked increase in bladder and lung cancer mortality in a region of Northern Chile due to arsenic in drinking water. *American Journal of Epidemiology*, 147(7):660–669.

[Stern and Cressie, 2000] Stern, H. S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19(1718):2377–2397.

[Swall and Davis, 2006] Swall, J. L. and Davis, J. M. (2006). A Bayesian statistical approach for the evaluation of CMAQ. *Atmospheric Environment*, 40(26):4883–4893.

[Tager et al., 2005] Tager, I. B., Balmes, J., Lurmann, F., Ngo, L., Alcorn, S., and Künzli, N. (2005). Chronic exposure to ambient ozone and lung function in young adults. *Epidemiology*, 16(6):751–759.

[Tong and Mauzerall, 2006] Tong, D. Q. and Mauzerall, D. L. (2006). Spatial variability of summertime tropospheric ozone over the continental United States: Implications of an evaluation of the CMAQ model. *Atmospheric Environment*, 40(17):3041–3056.

[Tonne et al., 2007] Tonne, C., Melly, S., Mittleman, M., Coull, B., Goldberg, R., and Schwartz, J. (2007). A case-control analysis of exposure to traffic and acute myocardial infarction. *Environmental Health Perspectives*, pages 53–57.

[Van Donkelaar et al., 2006] Van Donkelaar, A., Martin, R. V., and Park, R. J. (2006). Estimating ground-level PM2.5 using aerosol optical depth determined from satellite remote sensing. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 111(D21).

[Verbeke and Molenberghs, 2009] Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer.

[Wade et al., 2006] Wade, K. S., Mulholland, J. A., Marmur, A., Russell, A. G., Hartsell, B., Edgerton, E., Klein, M., Waller, L., Peel, J. L., and Tolbert, P. E. (2006). Effects of instrument precision and spatial variability on the assessment of the temporal variation of ambient air pollution in Atlanta, Georgia. *Journal of the Air & Waste Management Association*, 56(6):876–888.

[Wall, 2004] Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121(2):311–324.

[Waller et al., 1997] Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association*, 92(438):607–617.

[Wei et al., 2009] Wei, Y., Han, I.-K., Shao, M., Hu, M., Zhang, J., and Tang, X. (2009). PM2.5 constituents and oxidative DNA damage in humans. *Environmental Science & Technology*, 43(13):4757–4762.

[Weiss, 1997] Weiss, W. (1997). Cigarette Smoking and Lung Cancer Trends A Light at the End of the Tunnel? *CHEST Journal*, 111(5):1414–1416.

[Wikle, 2003] Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394.

[Wikle and Cressie, 1999] Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86(4):815–829.

[WONDER, 2014] WONDER (2014). United States Cancer Statistics: 1999 - 2011 Mortality, WONDER Online Database.

[Wong et al., 2004] Wong, D. W., Yuan, L., and Perlin, S. A. (2004). Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science and Environmental Epidemiology*, 14(5):404–415.

[Xia and Carlin, 1998] Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Statistics in Medicine*, 17(18):2025–2043.

[Yost et al., 2001] Yost, K., Perkins, C., Cohen, R., Morris, C., and Wright, W. (2001). Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. *Cancer Causes & Control*, 12(8):703–711.

[Zanobetti and Schwartz, 2005] Zanobetti, A. and Schwartz, J. (2005). The effect of particulate air pollution on emergency admissions for myocardial infarction: a multicity case-crossover analysis. *Environmental Health Perspectives*, pages 978–982.

[Zeger, 1988] Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75(4):621–629.

[Zeger et al., 1988] Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.

[Zhang et al., 2009] Zhang, H., Hoff, R. M., and Engel Cox, J. A. (2009). The relation between Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth and PM2.5 over the United States: a geographical comparison by US Environmental Protection Agency regions. *Journal of the Air & Waste Management Association*, 59(11):1358–1369.

[Zheng et al., 2009] Zheng, J., Shao, M., Che, W., Zhang, L., Zhong, L., Zhang, Y., and Streets, D. (2009). Speciated VOC emission inventory and spatial patterns of ozone formation potential in the Pearl River Delta, China. *Environmental Science & Technology*, 43(22):8580–8586.

[Zhu et al., 1999] Zhu, L., Carlin, B. P., English, P., and Scalf, R. (1999). Hierarchical modeling of spatio-temporally misaligned data: relating tra c density to pediatric asthma hospitalizations. *To appear Environmetrics*.

[Zhu et al., 2003] Zhu, L., Carlin, B. P., and Gelfand, A. E. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, 14(5):537–557.